RGB-D Scene Representations for Prosthetic Vision

David Feng

Submitted for the degree of
Doctor of Philosophy
at the
AUSTRALIAN NATIONAL UNIVERSITY
November 2017

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Signed .....

David Feng

November 25, 2017

Acknowledgements

This thesis would not have been possible without the help and support of a great many individuals. This is true for no-one more than the members of my PhD supervisory panel. I express my sincere gratitude to my primary supervisor and supervisory chair Assoc. Prof. Nick Barnes for his continuous support of my study, his unflappable good humour and patience, and his mentoring throughout the program. Nick's immense knowledge and invaluable guidance have been instrumental in my advancement through the program and the development of my research skills. I also thank Dr. Shaodi You for all his practical advice and useful comments on my research, and for staying up and providing assistance to his students during numerous paper deadlines. I sincerely thank Dr. Janine Walker for helping me with my writing, for sharing her considerable experience on running user studies, and for all the encouragement and guidance that she has gifted me throughout the program. I am also deeply grateful to Dr. Chris McCarthy for encouraging me to begin the program in the first place, and for all the helpful suggestions on research and paper writing over the years.

I would like to express my appreciation to all of the participants and research assistants who were involved in the user studies that were run as part of this thesis, as well as the students who helped annotate the dataset presented in this thesis. Their generous contribution of time and effort has assisted greatly with the evaluation of the work presented in this thesis.

I would like to acknowledge the staff at the Australian National University, National ICT Australia, and Data61 for facilitating a supportive and productive research environment. I also express my appreciation to Bionic Vision Australia for giving me the chance to be part of a project with such an immense scope and real-world relevance. This work was funded by the Australian Government Research Training Program Scholarship, and the NICTA top-up scholarship. Conference travel funding was provided by the Australian National University, and National ICT Australia.

I thank my fellow students and my friends for their solidarity and companionship, despite my increasing absence and gradual descent into hermitry as the end of program approached.

I offer my heartfelt gratitude to my parents for their wisdom and support throughout my life, and to my brother and Weiwei for their encouragement during this time.

Finally, I thank my late grandmother for inspiring in me a love of learning which has had a profound influence on my life and led me to this point. Her memory will be with me always.

RGB-D Scene Representations for Prosthetic Vision

by

David Feng

Submitted for the degree of
Doctor of Philosophy

Abstract

This thesis presents a new approach to scene representation for prosthetic vision. Structurally salient information from the scene is conveyed through the prosthetic vision display. Given the low resolution and dynamic range of the display, this enables robust identification and reliable interpretation of key structural features that are missed when using standard appearance-based scene representations. Specifically, two different types of salient structure are investigated: salient edge structure, for depiction of scene shape to the user; and salient object structure, for emulation of biological attention deployment when viewing a scene. This thesis proposes and evaluates novel computer vision algorithms for extracting salient edge and salient object structure from RGB-D input.

Extraction of salient edge structure from the scene is first investigated through low-level analysis of surface shape. Our approach is based on the observation that regions of irregular surface shape, such as the boundary between the wall and the floor, tend to be more informative of scene structure than uniformly shaped regions. We detect these surface irregularities through multi-scale analysis of iso-disparity contour orientations, providing a real time method that robustly identifies important scene structure. This approach is then extended by using a deep CNN to learn high level information for distinguishing salient edges from structural texture. A novel depth input encoding called the depth surface descriptor (DSD) is presented, which better captures scene geometry that corresponds to salient edges, improving

the learned model. These methods provide robust detection of salient edge structure in the scene.

The detection of salient object structure is first achieved by noting that salient objects often have contrasting shape from their surroundings. Contrasting shape in the depth image is captured through the proposed histogram of surface orientations (HOSO) feature. This feature is used to modulate depth and colour contrast in a saliency detection framework, improving the precision of saliency seed regions and through this the accuracy of the final detection. After this, a novel formulation of structural saliency is introduced based on the angular measure of local background enclosure (LBE). This formulation addresses fundamental limitations of depth contrast methods and is not reliant on foreground depth contrast in the scene. Saliency is instead measured through the degree to which a candidate patch exhibits foreground structure.

The effectiveness of the proposed approach is evaluated through both standard datasets as well as user studies that measure the contribution of structure-based representations. Our methods are found to more effectively measure salient structure in the scene than existing methods. Our approach results in improved performance compared to standard methods during practical use of an implant display.

Thesis Supervisor: Nick M. Barnes
Title: Associate Professor

List of Publications

This section lists the publications that were written as part of this thesis. For each publication, the corresponding thesis chapter is specified.

1. David Feng, and Chris McCarthy. "Enhancing scene structure in prosthetic vision using iso-disparity contour perturbance maps." In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5283-5286. 2013. (Chapter 3)

2. David Feng, Janine Walker, Nick Barnes, and Chris McCarthy. "A bi-modal visual representation can enhance orientation and mobility performance with less than 20 phosphenes." *Investigative Ophthalmology & Visual Science* 55, no. 13, pp. 1799-1799. 2014. (Chapter 8)

3. David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. "Local background enclosure for RGB-D salient object detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2343-2350. 2016. Spotlight Presentation. (Chapter 6)

4. David Feng, Nick Barnes, and Shaodi You. "DSD: Depth Structural Descriptor for Edge-Based Assistive Navigation." In *Proceedings of the IEEE International Conference on Computer Vision-Workshops*, pp. 1536-1544. 2017. (Chapter 4)

5. David Feng, Nick Barnes, and Shaodi You. "HOSO: Histogram Of Surface Orientation for RGB-D Salient Object Detection." Accepted in *Digital Image Computing: Techniques and Applications*. 2017. Oral Presentation. (Chapter 5)

Contents

Chapter 1

Introduction

When we open our eyes, we receive a tremendous amount of visual information. In order to manage the influx of visual stimuli, the human visual system rapidly directs attention towards areas in the scene that are more likely to be important. This attention direction occurs through a combination of pre-attentive and top-down processes, and enables the focus of limited available cognitive resources on the most relevant regions in the scene [109]. These regions are referred to as being *salient* to the human visual system, and are usually characterised by contrast with their surroundings [67]. Visual saliency detection plays an integral role in interpreting the environment when performing everyday tasks such as navigation, by quickly drawing attention to important scene components such as trip hazards and room boundaries [155].

Visual prostheses offer the potential to restore lost visual function to individuals with retinal disease [66]. Prosthetic vision devices normally convey incoming light intensity to the user, however current devices have low resolution and dynamic range [65]. For example, the prototype Bionic Vision Australia (BVA) retinal implant has 20 active electrodes, and implanted electrodes can only convey up to 10 different brightness levels to the user [66]. These display constraints can lead to difficulty in interpreting the content of the display, and, in particular, the biological attention deployment mechanisms of normally sighted individuals are not applicable in a prosthetic vision display [28]. This makes it easy to miss details in the environment

such as small or low-contrast trip hazards in front of the user. Figure 1-1 shows a simulation of a prosthetic vision display.

This thesis aims to enable improved scene perception from prosthetic vision devices, by applying computer vision techniques to detect structurally salient components within the scene that are relevant to tasks of everyday living. These detected components can then be emphasised in the final device display, drawing the user's attention and emulating the function of biological visual saliency detection. As a result, the user is provided with scene representations that are easier to interpret while ensuring relevant information is displayed, allowing for more effective use of the display when performing everyday tasks.

1.1 Prosthetic Vision Systems

Visual prostheses are devices that aim to convey vision-based information to users who are blind or have low vision. Prosthetic vision systems generally consist of three components: a camera for capturing a view of the environment, a processing unit to convert the camera image into a stimulation pattern, and a display for conveying the stimulation pattern to the user [65, 7]. The process of converting the camera image into the stimulation pattern is referred to as vision processing [10]. We refer to the stimulation pattern produced by a given vision processing method as a visual representation, or scene representation.

Prosthetic vision devices can be categorised according to the type of display. There are an increasingly large number of display technologies, which include retinal implants [65, 138, 7, 115, 50], vibro-tactile mats [76, 141], and tongue display units [8, 75]. Retinal implants are electrode arrays that are surgically implanted into the retina to elicit visual percepts, vibro-tactile mats are grids of vibrating motors worn on the body that form tactile displays, and tongue display units are electrical pulse generators that perform electro-tactile stimulation of the tongue through surface electrodes. These display devices share the common limitations of low resolution - on the order of tens or hundreds of display elements - and low dynamic range [65, 124, 141].

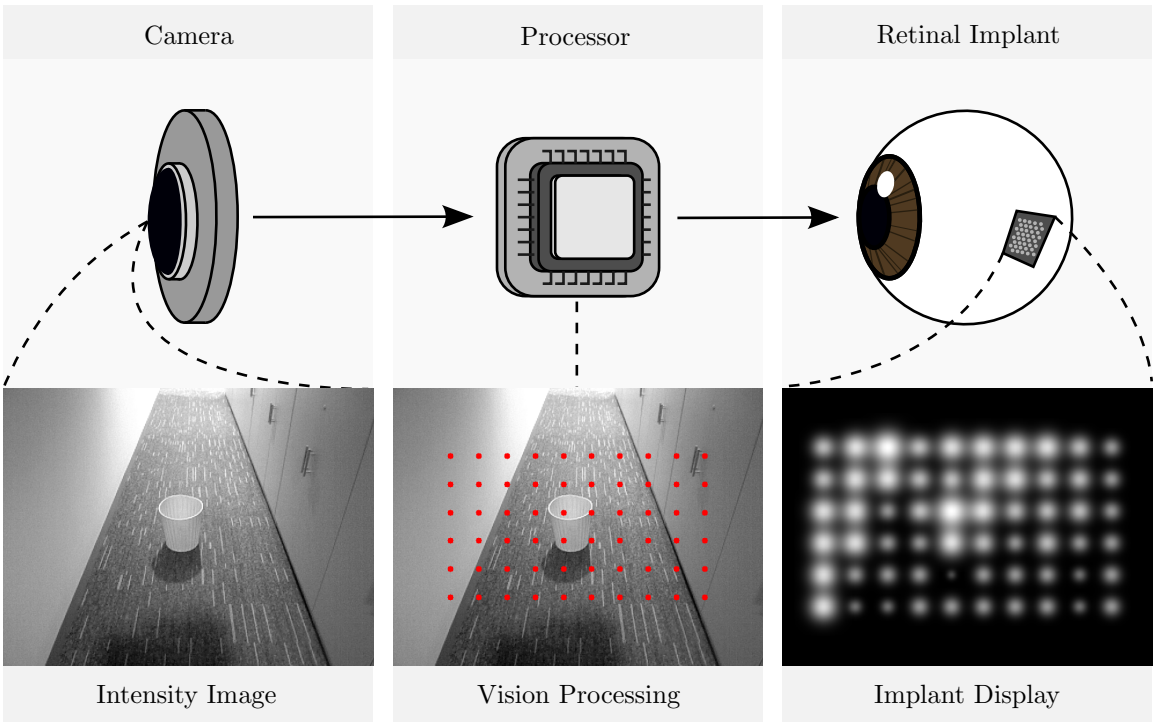


Figure 1-1. Overview of the major components and processing stages of a typical retinal implant device. The scene is captured through a body mounted camera. Vision processing is then applied to convert the camera image to the implant display image. In this example, the input intensity image is simply sampled at the output resolution, with sampling locations shown in red. Finally, the implant image is conveyed to the user through electrical stimulation of the retina. This example shows a simulation of what an implant user might see.

The work in this thesis is presented and evaluated in the context of retinal implant devices, although the proposed techniques are applicable to other prosthetic vision display types.

1.1.1 Retinal Implant Devices

Retinal prostheses have the potential to restore lost visual function caused by retinal dystrophies such as age-related macular degeneration and retinitis pigmentosa [65]. The progress of retinal implant development has been promising with a number of devices having been trialed in humans [65, 138, 7, 115, 50], and with two devices now available commercially. The Argus II [65] (Second Sight Medical Products) and the Alpha AMS [139] (Retinal Implant AG) have been granted the European CE

marking and/or FDA approval. While currently in its early stages, prosthetic vision has potential to improve the quality of life for many who have experienced vision loss from disease [128].

Retinal prostheses stimulate the human visual system using an implanted electrode array, bypassing non-functioning photoreceptors. Existing retinal implants are situated either near the inner surface of the retina (epi-retinal) [65, 115], underneath the retina (sub-retinal) [138], or behind the choroid (suprachoroidal) [7, 115, 123]. Stimulation is achieved by delivering a pulsed electrical current on each electrode. Upon activation of retinal neurons, implant users commonly report the perception of small spots of light in the visual field, referred to as phosphenes [38]. However, current prosthetic vision technologies are limited in terms of the number of implanted electrodes, with 1600 and 60 electrodes for the Alpha AMS and Argus II, respectively [65, 138]. Clinical studies show that implanted participants can distinguish up to ten different brightness levels using existing devices [66]. Thus, the resolution and dynamic range of electrically induced phosphenes are significantly lower than those of normal vision.

Current vision processing approaches for retinal implants predominantly employ intensity-based visual representations, in which stimulation levels convey the sampled light intensity near the projected electrode location in the visual field. However, the display constraints of near-term implants and the availability of high resolution input pixel data motivates the use of computer vision algorithms to ensure key task-relevant information is conveyed in the final display. Indeed, there is growing interest in the use of vision processing to boost functional outcomes with retinal prostheses [63, 163, 97, 12, 104], and demonstration that vision processing can lead to improved performance for retinal implant users [11].

1.2 Structural Saliency for Prosthetic Vision

Saliency detection methods have potential to further improve scene interpret-ability and associated functional outcomes for prosthetic vision devices. Saliency detection is the determination of what is most interesting in a scene [67], and in a prosthetic vision context serves to direct attention towards important parts of the scene.

We define *structural saliency* as the determination of what is important based on scene structure. The work presented in this thesis uses structural information obtained from depth images, in which pixel values correspond to distances between the camera and surfaces in the scene. To understand the motivation for using scene structure, we note that many everyday tasks, including navigation and grasping, involve interaction with the environment, and therefore require knowledge of physical quantities such as the locations and arrangements of the scene components relevant to the interaction. For example, picking up a hot mug requires knowledge of the location and orientation of the mug handle, while walking through a corridor requires knowledge of the orientation of the walls for selecting and maintaining an appropriate direction of travel. When computing structural saliency, these quantities are directly measured from the depth image, enabling the reliable extraction of regions exhibiting important structure that support improved perception on the prosthetic vision display.

1.2.1 Advantages of Structural Saliency

Previous work has largely focussed on appearance-based saliency computation from colour features, such as colour contrast, rather than structural saliency computation [21]. In particular, recent work has begun to examine the application of appearance-based saliency for prosthetic vision [112]. However, structural saliency provides a number of advantages over appearance-based methods for prosthetic vision.

Firstly, structure is more relevant than appearance at the level of perception afforded by the prosthesis display. Structural information enables direct measurement of relevant physical quantities for conveying scene shape and detecting potential haz-

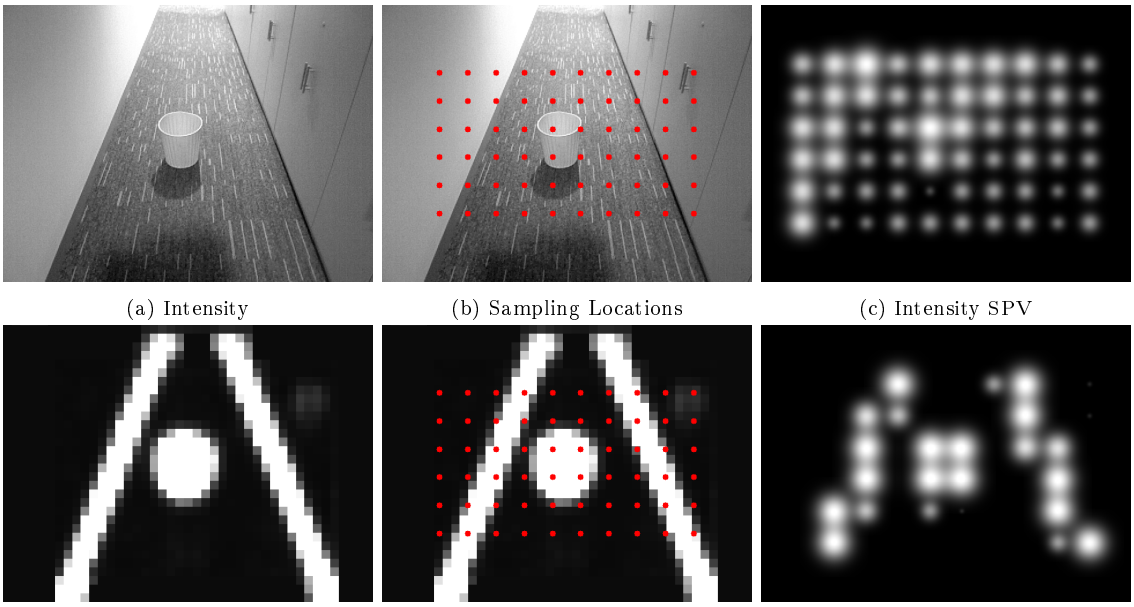
ards, whereas methods operating on scene colour may squander the limited prosthesis display space on information that is not relevant when performing physical tasks. For example, while a complex painting may be salient for a normally-sighted individual, it would be all but impossible to interpret within the limitations of the prosthesis display. On the other hand, scene structure such as corridor boundaries or the step of a sidewalk would be of immediate interest to a prosthesis user for understanding their surroundings while moving through the environment.

Furthermore, appearance-based methods suffer from false detections due to illusory edges and regions in the RGB image. One example of this is shadows, which are unavoidable in scenes containing objects and a light source displaced from the camera axis. Another example is texture, which occurs frequently both in man-made environments and in nature, for instance on printed wallpaper or the plumage of a bird. Structure-based methods are able to ignore these distractors and operate based on the underlying shape of the scene.

Appearance-based methods are also unreliable under low contrast conditions, for example when lighting is poor or when objects have a similar coloring to their surroundings. Depth capture techniques such as structured light and LIDAR are robust to lighting conditions and object colouring, allowing them to reliably perform detection under these conditions. The improved robustness offered by structural saliency compared to appearance-based saliency is particularly important for tasks such as navigation, where missing a trip hazard or the first step in a staircase can have disastrous results.

1.3 Research Problem

This thesis investigates structural saliency, and how knowledge of 3D scene structure can improve understanding of the scene in order to inform a prosthetic vision scene representation. The aim of the research presented is to detect structurally salient regions of the scene in order to ensure that relevant information is available within the capacity of the prosthetic vision display. The problem that the thesis aims to



(a) Intensity (b) Sampling Locations (c) Intensity SPV

(d) Structure (Salient Edges) (e) Sampling Locations (f) Structure SPV

Figure 1-2. Comparison of simulated prosthetic vision (SPV) of a scene based on intensity and structural salient edge detection. The simulation parameters approximate what a user might see when using the Argus II 60-electrode array from Second Sight medical products. Note that the intensity SPV image is difficult to interpret due to the low dynamic range and resolution of the implant display. Salient edge detection identifies and conveys boundaries that support interpretation of a scene, in this case the wall-floor boundaries and the obstacle, from depth information.

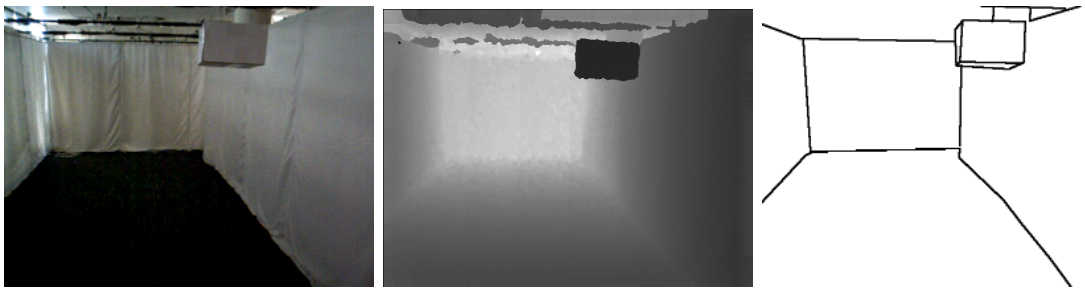
address is thus:

- How can information about the 3D structure of a scene inform a prosthetic vision display?

This thesis investigates the research problem by examining two subproblems, which are: salient edge detection; and, salient object detection.

1.3.1 Salient Edge Detection

Firstly, we note that physical tasks such as navigation and grasping are commonly considered an integral part of everyday living. These types of tasks generally require interaction with the environment. Therefore, visually guided performance of physical tasks requires extracting information about the arrangement of the scene components



(a) RGB (b) Depth (c) Salient Edges

Figure 1-3. Example of salient edges in a scene. Note that the salient edges clearly denote important boundaries for interpreting the shape of the scene, which would be useful for a task such as navigation.

relevant to the interaction. We refer to this as the "shape" of the scene. In normal human vision, scene shape is inferred through complex appearance-based visual cues such as colour [69], shading [118], and stereopsis [64]. However, none of these cues are able to be effectively conveyed due to the limitations of the display. In this situation, scene shape must be presented to the user in a more explicit manner. We propose to convey the shape of the scene by displaying significant structural boundaries to the user. We refer to these structurally significant boundaries as "salient edges". Clearly delineating the important structural boundaries in the scene enables the user to reliably interpret the shape of the scene, as shown in Figure 1-3. This leads to the first subproblem of the thesis:

- How can we model salient edge structure in order to convey scene shape through the prosthetic vision display?

1.3.2 Salient Object Detection

Secondly, we note that when first viewing a scene, biological vision systems rapidly fixate on certain salient parts of the scene, identifying the most important regions on which to further focus limited cognitive resources [109]. This process would be particularly useful in prosthetic vision where the display limitations severely restrict the amount of visual information available, and it thus becomes more important to

(a) RGB (b) Depth (c) Salient Object

Figure 1-4. Example of a salient object in a scene. Salient object detection can be used to direct visual attention in a prosthetic vision display to parts of the scene that stand out to normal human visual perception.

focus this limited visual perception towards salient regions rather than unimportant regions. However, biological attention direction does not occur in any meaningful way on a prosthetic vision display because the low-level visual cues that are necessary for saliency detection are not available within the display limitations. Therefore, we aim to substitute this biological process with computational modelling of the type of structure that is salient to normal human vision. These detected regions are commonly referred to as "salient objects" [19]. An example of a salient object is shown in Figure 1-4. The second subproblem of the thesis is:

- How can we model salient object structure for attention direction in a prosthetic vision display?

1.4 Approach

In order to address the research problem, we propose methods to extract salient edge and salient object structure from the scene, and evaluate the effectiveness of these methods.

We first investigate salient edge structure by detecting regions that have irregular surface shape compared to their surroundings. This is based on the observation that irregular surface regions support understanding of a scene. For example, areas with a largely consistent surface shape, such as a featureless wall, are not salient. On the other hand, the boundary between the ground and a wall for example has

irregular surface shape and therefore is salient. In order to measure this, we note that iso-disparity contours provide a rich description of surface shape, and thus the desired surface irregularities can be quantified through low-level iso-disparity contour analysis.

Our low-level model of surface irregularities is then extended to incorporate general scale and context information using a convolutional neural network. Scale and context information model the top-down processes that occur in biological saliency, and are necessary for obtaining accurate results given the variety of challenging scenes encountered during everyday living tasks [157]. For example, surface irregularities on curtain ripples should be less salient than on the boundary between two perpendicular curtains. We note that with dataset and algorithm limitations, careful encoding of the depth input is important for obtaining good results. We introduce a minimal encoding of the depth image based on surface orientation that gives better results than the state-of-the-art encoding on an edge detection dataset. In addition, our method is able to more accurately identify features that are important for understanding a scene on a custom prosthetic vision navigation dataset.

The second thesis subproblem is the detection of objects and regions in the depth image that are salient to the human visual system. In a prosthetic vision scenario, this type of saliency would be helpful for grasping objects or avoiding trip hazards. We propose to perform this detection based on the insight that salient regions tend to have contrasting surface shape from their surroundings. Detection is thus performed by combining surface orientation contrast and depth contrast in order to locate high-precision candidate regions in the image that have both different depth and shape from their surroundings. These seed regions are then expanded with post-processing in order to segment the salient object.

We then introduce a novel low-level depth saliency feature for salient object detection called Local Background Enclosure (LBE) that improves on contrast-based measures in existing depth saliency systems. Whereas contrast-based methods assign saliency according to the depth difference between an object and its surroundings, LBE measures the extent to which an object is in front of its surroundings. It is shown

that this formulation of low-level depth saliency produces better raw results as well as overall saliency system results compared to depth contrast. This provides insight into the types of structure that are salient to the human visual system, demonstrating that regions with background enclosure structure are more likely to be salient than regions exhibiting depth contrast.

While this thesis proposes many structural saliency techniques aimed at prosthetic vision scene representation, it remains to be shown whether these types of methods can be used effectively to perform tasks of everyday living with prosthetic vision, and whether they offer any advantage over standard scene representations during practical use of a prosthetic vision display. We verify the feasibility of using these types of exploratory scene representations through two prosthetic vision user studies.

First, we test the surface irregularities visual representation against the standard intensity-based downsampling method on an orientation and mobility task. The surface irregularities visual representation was found to result in a lower collision rate in a pilot study with participants using simulated prosthetic vision. A similar result was obtained in a subsequent clinical study with retinal implant users with the prototype BVA 24-electrode array implant [13]. This demonstrates that structural saliency methods provide relevant information that aids environment perception and improves navigation task performance compared to standard methods.

Second, we test a new scene representation that merges both intensity and structural information, in which a subset of the display is mapped to obstacle proximity and position. Unlike previous depth-only or intensity-only representations, this bi-modal representation allows the user to utilise intensity landmarks for orientation, while providing robust real-time obstacle information from the depth image for obstacle avoidance. Our bi-modal representation results in significantly fewer collisions than the standard scene representation on an orientation and mobility task with simulated prosthetic vision. These results demonstrate that users are able to interpret complex multimodal cues to effectively move through an environment, and shows that the inclusion of structural information significantly improves functional outcomes.

1.5 Major Contributions

In order to address the research problem, a number of major contributions are presented in this thesis:

- A new method that measures salient structure from surface irregularities is introduced. This method supports understanding of the scene on a prosthetic vision display by conveying scene structure.

- A method that improves the surface irregularities detection of salient structure is introduced, incorporating learned high level information to distinguish structurally salient edges.

- An investigation of local surface shape contrast to measure structurally salient object locations is performed. Our method enables improved localisation of salient objects based on their shape. This method enables vision processing methods to more closely model human visual saliency.

- A background enclosure-based formulation of depth saliency is proposed; this is a new approach that addresses intrinsic limitations of existing depth-contrast methods, better capturing scene structure that is salient to the human visual system.

- The effectiveness of the surface irregularities method for navigation with prosthetic vision is demonstrated in a pilot study involving prosthetic vision simulation.

- The first scene representation that merges colour and structural information for prosthetic vision has been developed and is evaluated in a simulated prosthetic vision user study.

- The work has been written as a series of publications in high-impact conferences. For more details please see page 9.

1.6 Thesis Map

This dissertation is organised as follows. Chapter 2 provides a literature survey. The proposed salient edge detection methods are proposed in Chapters 3, and 4, and the structural salient object detection methods in Chapters 5, and 6. User study evaluations are presented in Chapters 7 and 8. Chapter 9 concludes the thesis. See Figure 1-5 for a diagram of the major chapters of the dissertation. A detailed chapter list is given below.

- **Chapter 2** gives an overview of existing work in structural salient object detection, structural edge detection, and vision processing methods for prosthetic vision.

- **Chapter 3** describes the surface irregularities saliency system, which aims to convey scene structure to the user. This method has been designed specifically for scene representation for prosthetic vision.

- **Chapter 4** extends surface irregularities with learned high-level information. This allows the system to handle texture and improve predictions based on context, scale and other information. We use a deep CNN architecture with a novel depth image encoding.

- **Chapter 5** explores the role of surface orientation contrast for structural salient object detection. A new salient object detection system is developed, based on the idea that salient objects are likely to have different shape to their surroundings.

- **Chapter 6** proposes the local background enclosure feature, a new way of measuring structural saliency that more closely captures salient object structure compared to existing contrast-based methods.

- **Chapter 7** evaluates the surface irregularities scene representation from Chapter 3 in a user study with both simulated prosthetic vision and retinal implant

users. The method is shown to improve performance on a navigation task compared with the state-of-the-art scene representation.

- **Chapter 8** introduces a new scene representation that combines scene intensity information with structure-derived object cues. This multi-modal representation is tested in a simulated prosthetic vision user study on a navigation task, and is shown to improve performance compared with the standard scene representation.

- **Chapter 9** concludes this thesis, providing a summary of its contributions. The limitations of the work presented in the thesis as well as directions for future research in the area are also discussed.

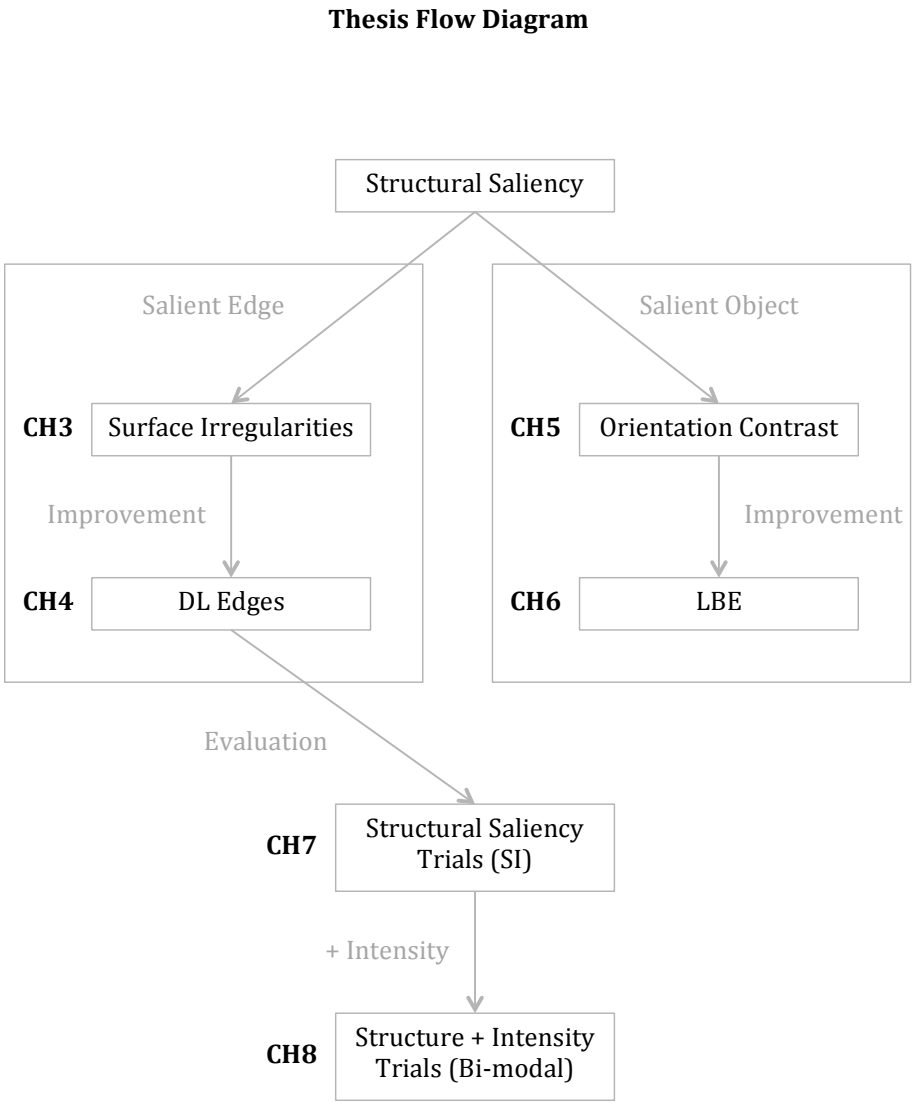


Figure 1-5. Organisation of main dissertation chapters.

Chapter 2

Background

This thesis aims to enable more effective prosthetic vision scene representations based on the measurement of salient structure in the scene. The work presented in this thesis falls across three main topic areas: structural salient object detection, structural edge detection, and vision processing for prosthetic vision. This chapter will present background and related work for each of these areas. First we introduce the notational conventions that will be used to present existing methods in this chapter as well as our proposed methods in the later chapters of the thesis. Then we will provide an overview of salient object detection, followed by structural edge detection, and finally we will review existing vision processing methods for scene representation with retinal implants. Note that in this chapter the ordering of structural salient object detection and structural edge detection is different compared to the order of the respective chapters in the dissertation. The topic arrangement in this chapter reflects the chronological order in which the areas were reviewed.

2.1 Notation and Definitions

The RGB-D input to our saliency detection methods is a single view of the scene, which consists of a colour image and a depth image. The colour image $I : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ maps from pixel positions to RGB colour, and the depth image $D : \mathbb{R}^2 \rightarrow \mathbb{R}$ maps the same set of pixel positions to scene depth, *i.e.* the distance between the image

plane and the 3D point corresponding to the pixel position. We assume that these two images are registered, such that $I(x, y)$ and $D(x, y)$ denote the colour and depth of the same location in the scene. We will refer to the set of possible pixel positions as $\text{dom}(D)$, *i.e.* the domain of the depth map.

When referring to an image patch or region, such as a superpixel obtained from image segmentation, we will use the notation $P \subset \mathbb{R}^2$ to denote the set of pixel positions that belong to the region. For notational convenience, we define commonly used properties of image patches as follows:

- Mean depth:
$$D(P) = \frac{1}{\text{card}(P)} \sum_{(x,y)\in P} D(x, y). \tag{2.1}$$

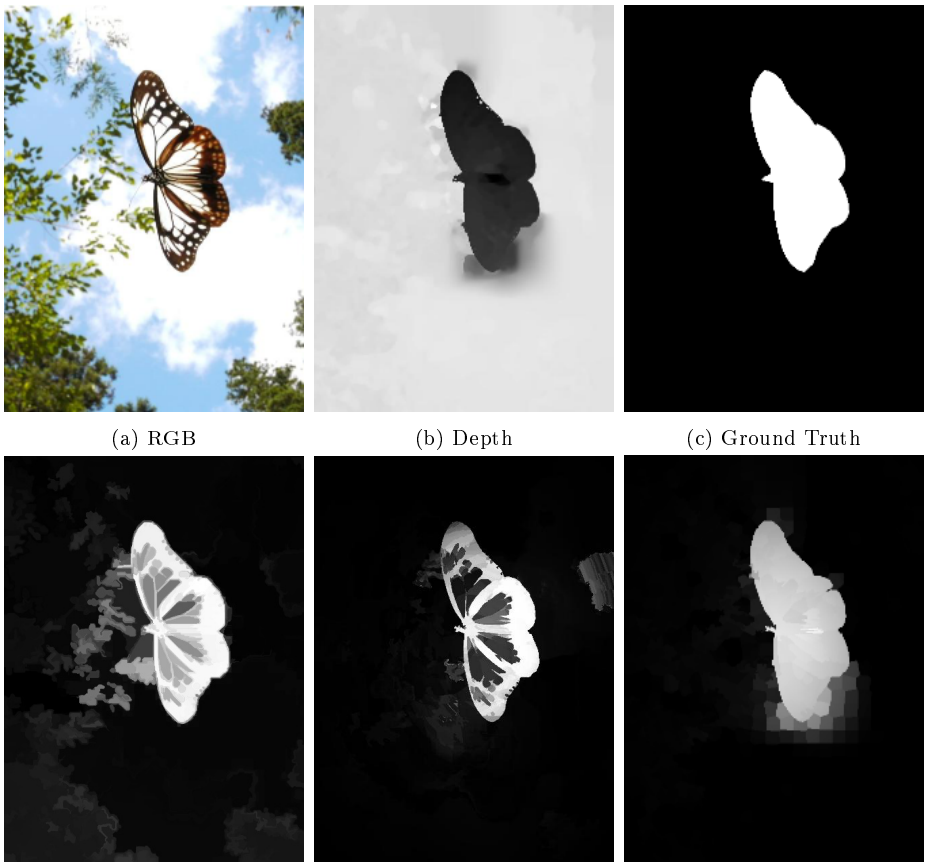
- Mean RGB colour:
$$I(P) = \frac{1}{\text{card}(P)} \sum_{(x,y)\in P} I(x, y). \tag{2.2}$$

- Centroid:
$$(x_P, y_P) = \frac{1}{\text{card}(P)} \sum_{(x,y)\in P} (x, y). \tag{2.3}$$

We will use $\mathcal{P}$ to denote to a set of patches that form a segmentation of the input image.

2.2 Structural Salient Object Detection

Visual attention refers to the ability of the human visual system to rapidly identify scene components that stand out, or are salient, with respect to their surroundings. Early work on computing saliency aimed to model and predict human gaze on images [68]. Recently the field has expanded to include the detection and segmentation of entire salient regions or objects [93]. This is referred to as salient object detection. Effective modelling of salient regions and objects can inform scene representations in a prosthetic vision scenario, enabling the cues provided by scene representations to more closely emulate biological attention direction.



(a) RGB (b) Depth (c) Ground Truth

(d) Saliency (DRFI) [70] (e) Saliency (DSR) [88] (f) Saliency (ACSD) [73]

Figure 2-1. Illustration of a salient object detection scenario, with saliency maps generated using appearance (d) and (e), and depth (f). Note that in this example depth information can help distinguish the salient object.

Salient object detection methods aim to locate and segment the regions in the scene that appear most interesting with respect to human perception [21]. Specifically, the goal of salient object detection is to identify the subset of pixels that are part of a salient region in the input image. Error is measured according to the salient object ground truth, a binary pixel map produced by human annotators, which denotes each pixel that is part of a salient object. An example image and the corresponding salient object ground truth are shown in Figure 2-1.

Existing methods typically produce a saliency score for each pixel that represents how likely the pixel is to be salient [68]. Note that saliency computation is not necessarily performed per pixel, as it is common practice to compute saliency scores

for homogeneous regions, such as superpixels, in the image and propagate the region scores to their member pixels [29]. The set of saliency scores produced by a method is referred to as the salient object map, or saliency map. An example of a saliency map from an existing method is shown in Figure 2-1. Due to the ground truth format, salient object detection methods must produce a saliency map that not only identifies but also accurately segments salient objects in the scene [21].

The saliency of a region is usually computed by measuring contrast at a local [68] and/or global scale [29]. The majority of previous approaches measure contrast with respect to appearance-based features such as colour, texture, and intensity edges. Structural information, however, can provide valuable cues for determining what is salient in the scene. In particular, structural information, such as depth images, enables object detection and attention prediction in a manner that is robust to appearance. For example, depth information can be used to suppress false detections from texture or shadows in the colour image as shown in Figure 2-1f, as well as supporting detection of objects under low contrast conditions. Structural saliency is driven by a wide variety of applications including robotic grasping [116], stereoscopic rendering [24], and scene understanding for assistive vision.

This section will first give a brief overview of salient object detection methods that operate on scene appearance. Following this, existing work on structural saliency will be presented, including depth in classical saliency systems, depth prior based methods, depth contrast based methods, and deep learning methods.

2.2.1 Salient Object Detection from Appearance

The majority of salient object detection methods operate on RGB input. Early interest in computational modeling of visual attention was sparked by the seminal work of Itti and Koch [68]. The major insight of this work was that areas of the image that exhibit high local center-surround contrast were more likely to be salient to the human visual system. An example is shown in Figure 2-1, where the salient object has high contrast with its local surroundings. Since then, colour contrast measurement has formed the foundation of many salient object detection methods

[4, 93, 80, 29, 30].

There are a wide variety of techniques for measuring contrast and detecting salient object regions in an RGB image. Achanta *et al.* [4] present a frequency-tuned model to detect salient objects, where contrast is measured as the pixel-wise difference between the average image colour and a Gaussian filtered image. Liu *et al.* [93] measure regional contrast through the Chi squared histogram difference between a rectangular image region and its surrounding region. Klein *et al.* [80] compute this quantity in an information theoretic way, using the Kullbeck Lebecker divergence metric. Cheng *et al.* [29] measure the global contrast between a superpixel and all other superpixels, taking into account spatial coherence. Cheng *et al.* [30] perform saliency computation from a soft abstraction of the image, allowing a larger spatial support and more uniform highlighting of objects compared to many superpixel based methods. Shen *et al.* [132] formulate saliency as a low-rank matrix recovery problem, where the background regions correspond to a low-rank matrix and the salient regions appear as sparse noise.

Prior knowledge about the task and human visual system plays an important role in salient object detection. The most widely used prior is the spatial prior, which performs re-weighting of the saliency score of a pixel according to the spatial distance between the pixel and the image center. This prior is used in almost all existing saliency systems, and is based on the biological tendency of the human visual system to focus on central image regions [146]. Similarly, many methods also make use of a background prior for saliency estimation, which exploits the idea that the borders of the image are more likely to contain background. Methods using the background prior compute saliency as contrast with the border region of the image, referred to as the 'pseudo-background'. Wei *et al.* [153] construct an undirected weighted graph where each superpixel and the pseudo-background are nodes, and saliency is computed as the geodesic distance between the superpixel and the pseudo-background. Li *et al.* [88] compute saliency as dense and sparse reconstruction errors with respect to the pseudo-background. Jiang *et al.* [70] use an absorbing Markov chain to compute saliency, in which border superpixels are absorbing nodes and non-border superpixels

are transient nodes. In this approach, the saliency of a superpixel is computed as the absorbed time from the transient node to the absorbing nodes.

Aside from spatial and background priors, many other heuristics have been used to improve salient object detection performance. For example, Liu *et al.* [93] observe that the spatial distribution of colour within an image correlates with saliency, since background colours are more likely to be spread out. Chang *et al.* [25] take advantage of the objectness prior, fusing the object proposal generation model of Alexe *et al.* [5] and region saliency detection in a graph-based framework.

Several approaches have employed machine learning methods to more closely model the properties that lead to an object being perceived as a salient. Liu *et al.* [93] learn a CRF model to segment salient objects based on a set of image features. Li *et al.* [87] use a SVM to predict saliency based on the difference between a target region and its local surroundings in feature space. Lu *et al.* [96] use a large margin framework to classify salient regions. While these methods offer good performance, the properties that make an object appear salient can be difficult to capture with linear classifiers. As such, subsequent approaches have used boosted decision trees [105], random forests [70], and a mix of linear SVMs [77] to measure saliency according to non-linear classification of regional descriptors.

Recently, deep learning methods have produced state-of-the-art results for salient object detection. Zhao *et al.* [162] propose a multi-contextual CNN for salient object detection, which jointly models local and global contexts of superpixels. This model is pretrained on the ImageNet dataset, due to the insufficient size of existing saliency datasets. Wang *et al.* [152] compute a local saliency map using a CNN and objectness based refinement, and then use a fully connected CNN to produce the final saliency map from global features of the object proposals. More recent methods have taken advantage of learned features from fully convolutional object detection networks such as VGG16 [135] and GoogleNet [142], which provide strong performance when fine-tuned for saliency detection [86, 92, 83]. Li and Yu [86] combine VGG16 with a region-based CNN to better model saliency discontinuities along object boundaries. Liu and Han [92] propose a hierarchical recurrent CNN based on VGG16, which predicts

saliency in a coarse-to-fine manner. Lee *et al.* [83] combine the high level features from VGG16 with a low level map which encodes distances between superpixel features. Fully convolutional object detection networks used in conjunction with optimisations for segmentation accuracy are the current state-of-the-art methods for appearance-based salient object detection.

2.2.2 Depth in Classic Saliency Systems

Compared to the large volume of work in appearance-based saliency computation, detection of salient object structure in a scene has been less explored. Early works in structural saliency use the raw depth image as an additional channel in the classic RGB saliency framework of Itti *et al.* [68]. Ouerhani and Hugli [110] explore which features to incorporate into the framework, selecting depth over the higher order properties of depth gradient and curvature. The authors note that the higher order features magnify noise in the depth image, reducing performance. Frintrop *et al.* [49] apply the framework to depth and intensity input in order to reduce the search space for object detection. However, saliency frameworks designed for RGB features are not ideal for structural analysis, since appearance-based features and depth-based features are two fundamentally different modes of representation. Therefore, applying the same heuristic frameworks to both types of features is an ad-hoc approach.

2.2.3 Depth Prior Methods

Based on findings that closer objects are more likely to appear salient in the human visual system [82], a number of techniques use depth values to modulate saliency maps computed from appearance. Zhang *et al.* [161] scales the output of [68] with depth to identify regions of interest in stereoscopic video. Similarly, Chamaret *et al.* [24] weights an RGB saliency map with depth values to identify salient regions for adaptive rendering on a 3D display. In these approaches, each pixel in the RGB saliency map is scaled by the corresponding depth, directly implementing the prior that closer regions tend to be more salient. Later saliency systems apply this depth

prior as a standard post-processing step, much like the spatial prior in appearance-based systems.

In addition to linear depth scaling, weighting RGB saliency based on a Gaussian distribution over depth has also been explored. Lin *et al.* [91] use a Gaussian distribution centered on the local maximum of a depth histogram to reweight an RGB saliency map. Tang *et al.* [144] attenuate saliency using a Gaussian computed from the depth values of salient regions, filtering object patches for salient object detection.

Some approaches aim to directly model the influence of depth on human visual attention by learning a non-linear depth prior from eye tracking data. Lang *et al.* [82] model the joint density between depth distribution and saliency response using a Gaussian mixture model learned from 3D eye tracking data, while Wang *et al.* [151] apply a learned mapping between saliency and difference of Gaussians response on the depth image. Depth prior based approaches generally do not consider relative depth, and work best when the foreground depth range is closer than the background.

2.2.4 Depth Contrast Methods

The effectiveness of global contrast for RGB salient object detection [29] has inspired similar approaches using depth. Niu *et al.* [108] extend [29] with disparity contrast for salient object detection in stereo image pairs. Fang *et al.* [46] measure global contrast over depth, colour, luminance, and texture to predict gaze in stereoscopic images. Peng *et al.* [113] compute saliency using depth and colour at both global and local scales. While the majority of previous work takes absolute depth differences when measuring depth contrast, some methods modulate depth contrast by the relative depths between regions. Cheng *et al.* [32] use global colour and depth contrast for salient object detection, with increased depth contrast from 'pop-out' regions. Ju *et al.* [73] compute saliency based on the average distance to minimum values encountered along a set of scanlines. This approach is sensitive to noise and the placement of the scan lines, which only provide a partial sample of the neighbourhood. Ren *et al.* [121] combine colour and depth contrast with an orientation prior to filter surfaces unlikely to belong to salient objects.

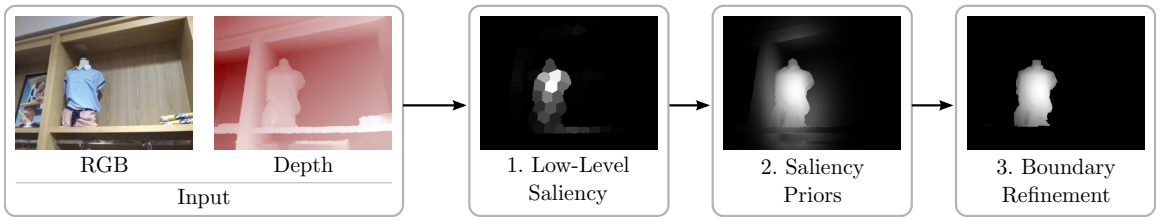


Figure 2-2. A typical RGB-D salient object detection pipeline. Existing methods compute a low level saliency map from depth contrast, followed by applying standard saliency priors, and finally performing boundary refinement.

Depth contrast based detection systems were the state-of-the-art methods in RGB-D salient object detection at the time the thesis contributions were made, and thus we will go over their common characteristics in detail. These methods are composed of a low-level contrast-based saliency detection operation, followed by various post-processing steps, as shown in Figure 2-2.

2.2.4.1 Low-level Saliency

State-of-the-art methods rely on depth contrast to locate salient regions [113, 121, 73]. In depth contrast-based methods, saliency is assigned according to the distance between the foreground and background. These methods perform saliency detection on an over-segmentation of the input image into superpixels, and later propagate superpixel saliency values to their member pixels [2]. Superpixels represent perceptually homogenous regions in the image, and thus are a useful abstraction that greatly increases computation speed. Furthermore, oversegmentations of the image retain the relevant object boundaries, supporting accurate object segmentation. In the following, we use P to denote an arbitrary superpixel in the input image.

Low-medium-high (LMH): Peng *et al.* [113] propose the LMH system, which assigns the saliency of P as the product of the contrast between P and three different contexts

$$S_{\text{LMH}}(P) = \prod_{N \in \{N_L, N_B, N_G\}} C_{\text{LMH}}(P, N), \tag{2.4}$$

where N_L is the local context which consists of the 32 closest patches to P, N_G is the global context which is the set of all patches, N_B is the pseudo-background

context which consists of the 36 patches closest to the image corners, and C_{LMH} is the contrast function. The contrast function uses kernel density estimation to estimate the probability that P belongs to a given context.

$$C_{\mathrm{LMH}}\left(P,N\right)=-\log\left(\frac{1}{\mathrm{card}(N)}\sum_{Q\in N}\mathrm{card(Q)}\exp\left(-\frac{\left(D(P)-D(Q)\right)^{2}}{2\sigma_{N}^{2}}\right)\right), \tag{2.5}$$

where σ_N is the bandwidth of the context N.

Global priors (GP): Ren *et al.* [121] propose the GP system, which is based on the global contrast saliency measure from [31]. The GP system computes the saliency of a superpixel as its weighted depth contrast with all other superpixels in the image.

$$S_{\mathrm{GP}}\left(P\right)=\sum_{Q\in\mathcal{P},Q\neq P}\mathrm{card}(Q)C_{\mathrm{GP}}\left(P,Q\right), \tag{2.6}$$

where the contrast measurement function C_{GP} is given by

$$C_{\mathrm{GP}}\left(P,Q\right)=\exp\left(-\frac{\|(x_{P},y_{P})-(x_{Q},y_{Q})\|}{2{\sigma_{\mathrm{xy}}}^{2}}\right)|D(P)-D(Q)|, \tag{2.7}$$

where σ_{xy} is the standard deviation of the distance between two region centroids.

Anisotropic center-surround difference (ACSD): Ju *et al.* [73] propose ACSD, based on the observation that while a salient object should be in front of its surrounds, patches on that object may be at a similar depth. Saliency is computed by placing scanlines over the patch and averaging the maximum distance to a patch along each line.

$$S_{\mathrm{ACSD}}(P)=\sum_{Q\in N(P)}B_{\mathrm{ACSD}}(Q)\left(D(P)-D(Q)\right) \tag{2.8}$$

where $N(P)$ is the directional neighbourhood of P, and B_{ACSD} is the background selection function given by

$$B_{\mathrm{ACSD}}(P,Q)=\begin{cases}1, & \text{if } D(P)\geq D(Q).\\ 0, & \text{otherwise}\end{cases} \tag{2.9}$$

The directional neighbourhood $N(P)$ is defined as the superpixels that intersect any of eight equally spaced lines originating from the centroid of P.

These methods use depth contrast as the fundamental saliency term, such that saliency reflects the depth difference between the foreground and the background. Depth contrast methods are unlikely to produce good results when a salient object has low depth contrast compared to the rest of the scene. Furthermore, depth contrast methods are generally unable to distinguish salient objects that have a different local surface shape to their surroundings.

2.2.4.2 Saliency Refinement

Salient object detection performance can be highly dependent on the accurate delineation of the object contour during segmentation. While depth contrast measurement forms the foundation of many approaches, it focuses on identifying salient object regions rather than on the object boundaries. This produces saliency maps with sparse saliency activations, or boundaries that are blurry and not well-defined. Thus, it is common practice to enhance these low-level saliency maps by applying priors and other refinement steps [21].

Saliency priors are ubiquitous in saliency detection, and involve adjusting the saliency value of a pixel or region based on prior knowledge. The most widely used priors for structural saliency detection are the spatial prior, which increases the saliency score of regions close to the image center, and the depth prior, which increases the saliency score of regions close to the observer. The use of these priors is widespread in existing work [113, 73, 32, 57, 136]. Ren *et al.* [121] explore orientation and background priors for detecting salient objects. The orientation prior increases the saliency score of regions parallel to the camera plane, whereas the background prior lowers the saliency of regions that have similar appearance to the image boundary.

In addition to priors, salient object detection systems use several other techniques to improve the segmentation. Peng *et al.* [113] perform object grouping in order to infer object boundaries from a sparse low level saliency detection output, by increasing the saliency score of regions with similar depth and image position to detected high-

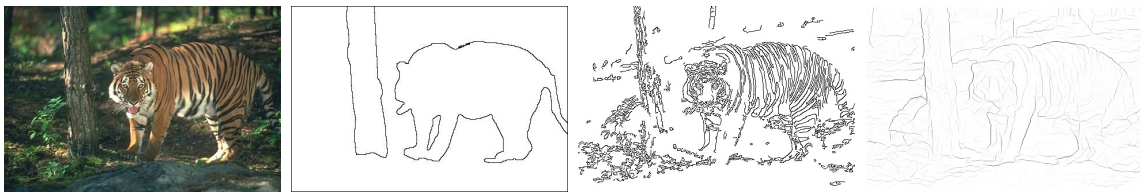
saliency seed regions. The output object saliency map is further refined with a region growing approach. Ju *et al.* [73] apply Grabcut segmentation to refine the boundaries of the low-level saliency map, improving object boundary delineation based on appearance. Ren *et al.* [121] use PageRank to improve salient object detection, and enforce spatial consistency of saliency labels using a Markov random field. Guo *et al.* [56] segment salient objects by using cellular automata to iteratively propagate saliency values from the low-level saliency map based on object boundary information.

While these post-processing steps significantly improve salient object detection results, the fundamental basis of saliency is the low-level saliency detection method itself, since it is through this method that the salient regions are initially identified. However, in the literature there has been relatively little development of the low-level saliency term compared to the post-processing steps, with existing methods relying on depth contrast to find salient regions.

2.2.5 Deep Learning

The work described in this section was published after the main contributions of the thesis, and is presented here for completeness.

Deep learning methods have recently produced state-of-the-art results for RGB-D salient object detection [27, 117, 133]. These techniques learn high level context and scale information that can greatly improve the identification of salient object structure. A major challenge with applying deep learning methods for structural salient object detection is the relatively small size of available RGB-D datasets [113, 73], which are an order of magnitude smaller than many commonly used RGB datasets [29, 147, 158]. Furthermore, state-of-the-art appearance-based salient object detection systems commonly fine tune networks trained for object detection from RGB data, greatly reducing the number of training examples required compared to training from scratch [162, 84]. However, there do not exist any corresponding pre-trained models for depth data. In order to address this issue, Chen *et al.* [27] apply cross-modal transfer learning to train a network for depth salient object detection using the weights from an existing RGB model for supervision. Qu *et al.* [117] take a different approach,



(a) RGB (b) Ground Truth (c) Canny [22] (d) Ren & Bo [156]

Figure 2-3. Illustration of a challenging contour detection scenario. Generated edge maps are included from (c) a low level edge detector [22], and (d) a contour detector [156].

performing training from scratch but on a set of salient features rather than the image pixel values. This greatly reduces training requirements, since the network learns a fusion of features rather than a mapping from raw pixel values. Shigematsu *et al.* [133] take a similar approach, using a CNN to learn a low-level distance map from low and mid-level depth features.

2.3 Structural Edge Detection

Edges are the boundaries between perceptual regions in a scene, and are ubiquitous in human visual perception. Edge detection mechanisms in the human visual system contribute to several low level visual processes, such as adjustment of object brightness across an edge, and induction of object contours from visible edges [130]. Edge detection is thus an important biological process that is integral to visual perception, and provides important cues for understanding one's surroundings.

There has been a large amount of interest in the computational detection of edges in computer vision. This is the problem of identifying the set of pixels that correspond to boundaries or object contours in an input image. Edge detection methods typically output a score for each pixel that reflects the likelihood of the pixel belonging to an edge, and the image containing the edge scores of all pixels is referred to as the edge map. Performance is evaluated according to ground truth binary edge maps, which denote the set of edge pixels as selected by a human annotator. Figure 2-3 shows an example image, the corresponding edge ground truth, and output edge maps produced by two edge detection methods.

The vast majority of existing work operates on scene appearance, identifying boundaries between regions based on intensity or colour. Early work typically used low level image filtering operations to detect all non-noise brightness discontinuities within a scene [22]. More recent approaches have aimed to reduce noise from texture by focusing on closed contour detection, using high level information to identify object contours and suppress edges within objects [6]. Example output from a low level edge detector and contour detector are shown in Figures 2-3c and 2-3d respectively.

Definition Structural edge detection is the task of finding edges between significant surfaces based on scene structure, *e.g.* from depth images. Rather than detecting edges from brightness discontinuities as in appearance-based edge detection, structural edge detection aims to detect boundaries based on surface shape. Compared to appearance-based methods, structural edge detection supports performing physical tasks using assistive vision since structural edge arrangements reflect scene structure. Detecting edges based on structure also offers advantages such as robustness to lighting conditions and texture.

This section provides an overview of existing work in edge detection, with a focus on methods that detect edges from structural input. First, background on low level edge detection methods is presented, followed by more recent work on contour detection.

2.3.1 Low-level Edge Detection

Early edge detection methods generally applied low level image filters to directly detect edges from local appearance, including classical approaches such as Sobel [78] and the highly successful Canny detector [22]. Classical work by Lowe *et al.* [95] linked concepts of perceptual organization to algorithms for finding lines, in order to understand 3D structure for object recognition. Extending on this, Ullman *et al.* [149] found 'structural saliency' (objects/regions) based on curvature, and Guy and Medioni [61] inferred probable boundary directions based on extending strong edges considering weak edges using a vote-based approach. These low level methods provide a high response on all strong edges regardless of semantic validity.

2.3.1.1 Classical Range Image Edge Detection

The 1990s saw an early wave of interest in range image edge detection for segmentation. Yokoya and Levine [160] compute edge maps for range image segmentation based on Gaussian and mean curvatures computed from a local biquadratic surface fit. This approach can be sensitive to sensor noise and quantization effects, and does not scale well to current depth capture resolutions.

Jiang and Bunke [72] perform edge detection on range image input using a scanline approximation approach. Each row, column, and diagonal of the range image is split into a set of polylines, and each polyline is fitted with a quadratic curve. Edge point locations are inferred from boundaries between curves, and contour closing is performed with morphological operators. Sappa *et al.* [126] reduce the effect of noise in this process by filtering small polylines, and also only consider rows and columns since two orthogonal directions are sufficient for scanline analysis. The dilation for contour closure is refined by enforcing dilation only along the direction of the line at endpoints, allowing the method to retain small lines that would have been lost to a general dilation operation. Sappa *et al.* [125] further refine the contour closure process by replacing the morphological operations with a graph partitioning approach, which can better handle many pathological open contour scenarios. However, the process of recursively fitting quadratic curves has a high computational cost, and accuracy also relies on setting an appropriate polyline splitting condition, which may be different for different images.

Bellon *et al.* [14] calculate a simple range image edge map to inform a clustering-based segmentation algorithm. The edge map is computed by adding a jump edge map to a roof edge map, obtained by comparing pointwise angle differences between adjacent normals. The authors subsequently refine the process using a skeletonisation method to perform thinning of the roof edge map, and remove small non-significant elements [15]. Pointwise comparison of normals can lead to highly noisy measurements from real world depth data from commodity depth sensors, since the significant noise in the depth image is magnified in its first order properties.

Choi *et al.* [33] detect high curvature edges for point cloud registration by applying a Canny edge detector to the surface normal image. However, careful tuning of the canny parameters is required to reduce errors from surface normal noise. Furthermore, large scale structural edges in the scene with low pointwise normal change may be missed.

2.3.2 Contour Detection

More recent work has found closed contours of objects using combined edge and region methods. For example, Levinshtein *et al.* [85] optimally grouped superpixels to find enclosing salient contours. Dollar and Zitnick [43] show strong results for detecting salient edges by learning using a structured forest and manually designed features. These are sometimes extended to RGB-D data. Ren and Bo [156] train sparse code gradients to detect contours, showing a significant performance boost in performance when adding depth data to RGB. Similarly, Dollar and Zitnick [43] show improved performance incorporating depth data when learning structured forest to perform edge detection. Raskar *et al.* and Schäfer *et al.* [127] explicitly use depth information to suppress texture intensity edges by requiring co-occurrence between depth and intensity edges [119, 127].

2.3.2.1 Deep Learning Methods

The excellent results yielded by CNNs for high level vision tasks such as object detection have led to revisiting contour detection. Early CNN contour detection approaches include Ganin and Lempitsky [51], Kivinen *et al.* [79], and Shen *et al.* [131]. State-of-the-art results have come from papers that transfer deep learning features from high-level vision tasks to low-level vision problems, including edge detection. Both Bertasius *et al.* [16], and Xie and Tu [157] derived contour detection from a base of VGGNet [135]. Using a pre-trained, trimmed VGGNet, [157] incorporate deep supervision to enforce meaningful output from intermediate layers as well as the final layer of the fully convolutional network. This has become the baseline model

for deep edge detection, with many subsequent papers proposing improvements to the architecture. Liu and Lew [94] propose relaxed deep supervision, using the output of off-the-shelf edge detectors to guide the learning of intermediate layers in a coarse-to-fine-paradigm. Kokkinos [81] fine tunes the loss function and explicitly incorporates multiple scales as well as global information. Maninis *et al.* [98] include a novel sparse boundary representation for hierarchical segmentation, and show that learning boundary strength and orientation improves results. Yang *et al.* [159] learn to detect contours with a fully convolutional encoder decoder network, which generalizes well to unseen object categories. These methods focus on RGB edge detection, and in particular do not investigate more effective representations of the depth data to improve detection of structural edges.

Depth Image Encoding In the context of object detection and scene segmentation, Gupta *et al.* [60] propose the HHA geocentric embedding for depth images to perform RGB-D contour detection. The HHA embedding encodes disparity, height above ground, and angle with gravity into the edge learning framework of [43] and show that it produces improved results over naively using depth. The HHA feature is the current state-of-the-art depth representation for CNN-based edge detection, with subsequent CNN-based edge detectors all incorporating this feature when operating on RGB-D input [157, 16, 131].

2.4 Vision Processing for Retinal Prostheses

Prosthetic vision displays are characterised by low resolution and dynamic range. Existing retinal implants contain between 16 [41] and 1550 [138] electrodes, which are able to convey up to ten different perceivable stimulation levels each [66]. The amount of information that an implant is able to convey is therefore orders of magnitude lower than provided by the cameras used for scene capture. For example, the Asus Xtion Pro camera performs colour and depth scene capture at a resolution of 640×480 pixels, with a dynamic range of 256 levels per colour channel, and 2048 levels for the depth channel. Vision processing is therefore required to convert the high resolution and

Figure 2-4. A mobile prosthetic vision simulation system, which consists of a head mounted RGB-D camera, head mounted display, and a processing unit worn on the back. Prosthetic vision simulation with normally sighted participants is commonly used to estimate performance while avoiding the overhead of running a clinical trial with implant users.

dynamic range camera image into a low resolution visual representation for display on the implant.

Due to the large amount of image information lost when reducing the resolution and dynamic range of the input data, the vision processing method is crucial in determining whether key task related information from the input image remains available in the prosthetic vision display. For instance, a vision processing method could preserve obstacle locations in the display for navigation, or detect and enlarge faces for the task of facial recognition. Vision processing methods that effectively convey task related information can improve functional performance on a task [111, 104].

The effectiveness of vision processing methods is evaluated through user studies, which measure the capacity of a scene representation to facilitate visually guided performance of a given type of task. Accurate estimation of vision processing performance requires testing in clinical trials, in which retinal prosthesis users make use of the visual representation to perform a task. However, due to the high overhead of performing a clinical trial, the majority of existing vision processing results are reported using simulated prosthetic vision (SPV) studies, which measure the performance of

normally-sighted participants completing tasks using an externally displayed rendering of phosphenes. Examples of SPV displays range from computer monitors showing renderings of pre-recorded or virtual environments [39], to mobile head mounted displays that convey the surroundings captured from a body worn camera [104], as shown in Figure 2-4. An example SPV image is shown in Figure 2-6c. SPV studies are useful for performing early evaluation of exploratory methods, helping to identify limitations and determine promising strategies on which to focus further resources such as clinical trials.

This section will provide an overview of existing vision processing methods for retinal implant systems. The work in this thesis is presented largely in the context of facilitating safe orientation and mobility, which is considered a highly desirable functional outcome for visual prostheses [23, 39]. Therefore this section will first provide a definition and overview of standard vision processing methods, and subsequently focus on related methods for orientation and mobility. Note that most existing methods are evaluated using SPV, which has several limitations. This section is thus concluded with an overview of the limitations of SPV.

2.4.1 Definition

Vision processing is the conversion of high resolution and high dynamic range sensor data into a low-resolution form for electrode stimulation. More precisely, Barnes *et al.* [10] formulate vision processing as the mapping

$$\Phi = \mathcal{F}(\Psi) \tag{2.10}$$

that converts from an input stream Ψ to a set of output phosphenes Φ, as illustrated in Figure 2-5. We assume the following constraints, which reflect the configuration of existing vision processing systems:

- Both the input and output are produced discretely, although not necessarily synchronously, with respect to time. Each stream can thus be split into frames.

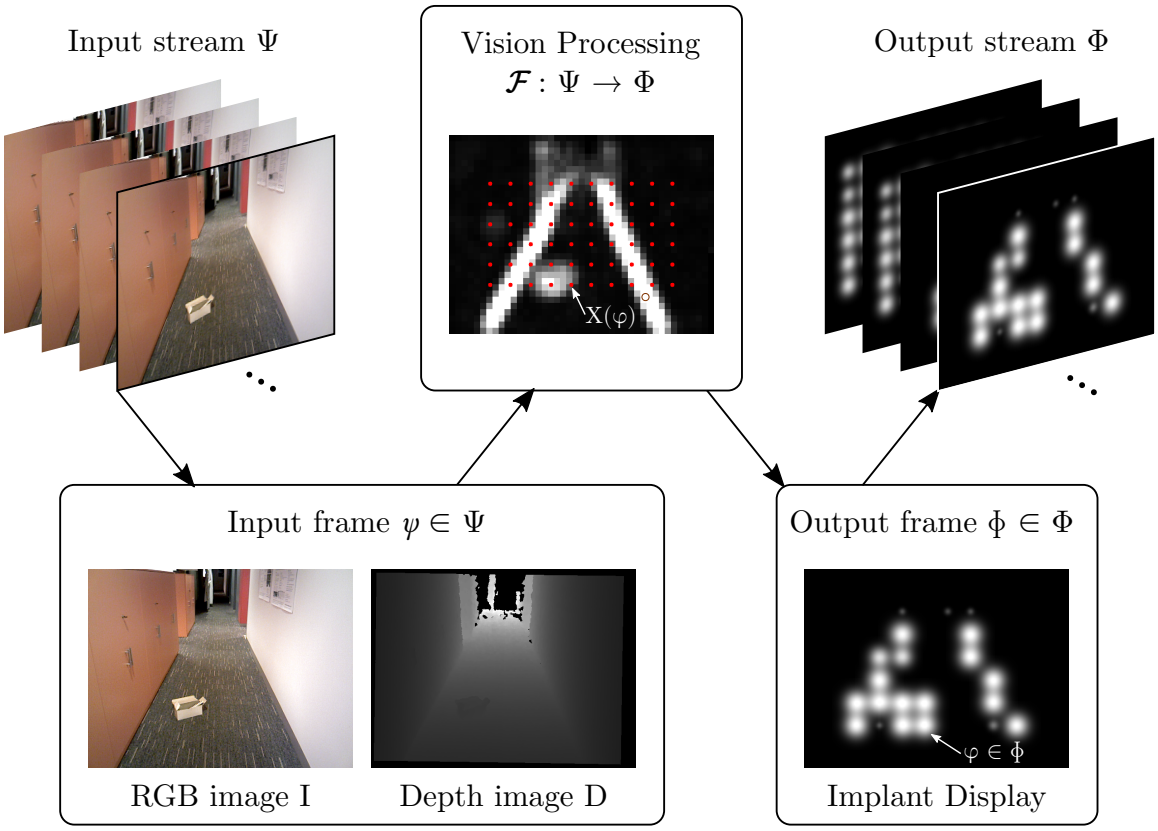


Figure 2-5. Illustration of our vision processing framework for prosthetic vision. Each input frame $\psi \in \Psi$ consists of an RGB image I and depth image D. The corresponding output frame $\phi \in \Phi$ is obtained by computing the stimulation level, or brightness, of each phosphene $\varphi \in \phi$. Note that we consider individual input frames separately.

- Mapping is performed on a per-frame basis, therefore every output frame $\phi \in \Phi$ is computed from a single input frame $\psi \in \Psi$.
- Each input frame ψ is composed of an intensity image I and/or a depth image D.

Under these assumptions, in order to specify $\mathcal{F}$ it is sufficient to define a phosphene brightness function $B(\varphi, I, D)$, which determines the output level for individual phosphenes $\varphi \in \phi$. In cases where there is only intensity or depth input, the brightness function will be $B(\varphi, I)$ and $B(\varphi, D)$ respectively.

It is common for $B(\varphi, I, D)$ to reflect the input image values at the projected location of φ in I and D. In the following sections we use $X(\varphi)$ to denote the projection of φ onto the visual field.

2.4.2 Standard Intensity Vision Processing

Current approaches predominantly employ intensity-based vision processing, in which stimulation levels convey the sampled light intensity near the projected electrode location in the visual field. This can be thought of as directly downsampling the input intensity image to the output display units. Although the implementation details of different approaches may differ slightly, this type of method will be referred to as the standard intensity method. An example of the standard intensity vision processing is shown in Figure 2-6.

The aim of these downsampling based vision processing methods is to filter all high frequency data, such as fine image details, noise, or texture, above a given cutoff without affecting lower frequencies, such as relatively large contiguous regions. Selection of an appropriate downsampling filter is important because if not all high frequency signals are filtered then aliasing can occur. An example of this effect in prosthetic vision would be if small changes in intensity on a single surface, e.g. from texture, cause significant and potentially misleading variations in phosphene brightness. Furthermore, some filters may also affect frequencies below the cutoff, which can result in a significant reduction in the sharpness of the filtered image.

In the standard intensity representation, the brightness $B(\varphi, I)$ of each phosphene φ is obtained by sampling the filtered input image at the projected location of the electrode $X(\varphi)$. Thus, the general phosphene brightness function for filtering based methods is given by:

$$B(\varphi, I) = (I * F) \circ X(\varphi), \tag{2.11}$$

where $I * F$ denotes the convolution of I with filter kernel F.

The minimal amount of vision processing is performed by simply setting phosphene brightness based on the intensity of the closest pixel to the projected phosphene location in the raw input image [23]. This is a special case of Equation 2.11 where $I * F = I$. This approach does not filter high frequency data within the input image and is prone to large output level variations caused by noise or texture. More recent methods use non-trivial filters, in order to ensure that the stimulation level of each

electrode is a more stable and representative depiction of the corresponding region of the visual field. Humayun *et al.* [65] use block filtering, averaging the intensity values within a block of pixels centred on the target pixel, setting F to a matrix of ones divided by the size of the filter. This type of filter has a high impact on frequencies below the cutoff, heavily blurring boundaries in the scene. Hayes *et al.* [62] set F to a Gaussian kernel, improving boundary sharpness by giving higher weights to pixels closer to the projected phosphene location. Dowling *et al.* [45] use median filtering, further improving the sharpness of edges in the input image.

2.4.2.1 Lanczos2 Filtering

In signal processing it is well-understood that Nyquist band-limited filtering prevents aliasing when performing image downsampling [154]. Lanczos2 offers a better compromise than other practical filters in reducing aliasing and retaining sharpness [148]. Barnes *et al.* [11] have shown that applying the Lanczos2 filter to downsample the input image at the Nyquist frequency is effective for prosthetic vision. Since sampling frequency is the inverse of sample spacing, the Nyquist frequency of a phosphene display can be estimated as $1/r$, where r is the average nearest-neighbour distance of each phosphene. Note that in a regular phosphene grid, r is equal to the phosphene spacing. This gives the 2D Lanczos2 reconstruction kernel of size $2r + 1$:

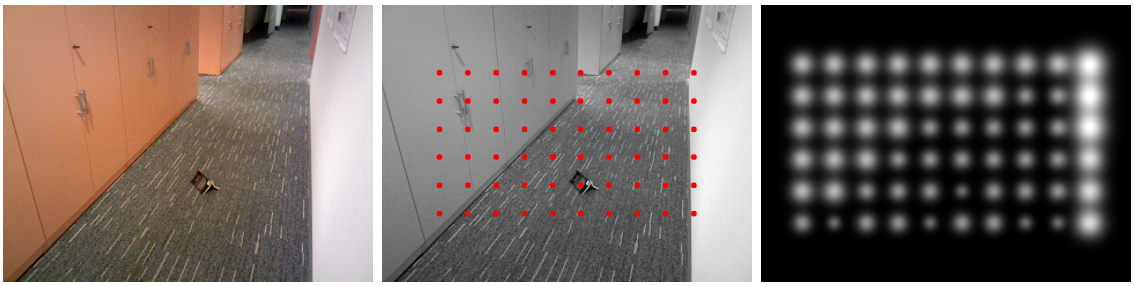
$$L = k\left(\frac{c(x, y)}{r}\right), \tag{2.12}$$

where $c(x, y) = \|(x, y) - (r, r)\|$ is the distance to the kernel centre, and k is the Lanczos2 kernel given by:

$$k(a) = \begin{cases} \text{sinc}(a)\text{sinc}\left(\frac{a}{2}\right), & \text{if } a \leq 2 \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

Therefore the vision processing method defined by

$$B_{\text{lanczos2}}(\varphi, I) = (I * L) \circ X(\varphi) \tag{2.14}$$



(a) RGB (b) Sampling Locations (c) Phosphene Image

Figure 2-6. Example SPV of a scene with the standard intensity vision processing method. Image (c) shows a simulation of what an implant user might be expected to see when viewing the scene using this visual representation with a 20-electrode retinal implant. Note the difficulty of interpreting this scene with the standard representation.

is the current state-of-the-art standard intensity method.

2.4.2.2 Clinical Trials

The standard intensity visual representation has been evaluated on an orientation and mobility task by Second Sight Medical Products LLC [65]. Using the Argus II epiretinal implant, participants performed the task of walking towards and touching a door-sized black target in a room with white featureless walls. Use of the implant with the standard intensity visual representation (system-ON) was found to result in a higher success rate than the control condition (system-OFF) in which no visual information was conveyed through the implant (55% system-ON vs. 31% system-OFF at 3 months with $N = 29$ participants; 60% system-ON vs. 8% system-OFF at 24 months with $N = 8$ participants). This demonstrates that the standard intensity representation can enable basic wayfinding for implant users assuming an environment with appropriate contrast.

2.4.2.3 Limitations

The standard downsampling-based vision processing methods aim to directly depict the scene intensity values from regular intervals of the input image, without assessing the importance of any part of the scene. However, this often means that details useful for the task, such as the top stair in a flight of steps, or the trip hazard

resting on the ground in Figure 2-6c, can be missed due to a lack of display capacity. Additionally, the standard intensity method relies on different scene components to have high contrast with each other in order to be discernible on the prosthetic vision display. Therefore it can be desirable to prioritise and enhance the display of crucial scene components.

2.4.3 Cueing Vision Processing Methods

There is growing interest in the application of computer vision methods to address the limitations of standard vision processing methods and boost functional outcomes with visual prostheses [112, 103]. These methods provide additional cues through the phosphene display to convey important information to the user.

2.4.3.1 On-Demand Saliency Cueing

Parikh *et al.* [112] provide on-demand peripheral object cueing derived from the output of a visual saliency algorithm [68]. While navigating using a standard downsampled intensity visual representation, the user may request a saliency cue, and receives the direction to a salient interest point via eight cueing phosphenes placed on the border of the screen. The inclusion of the peripheral cueing phosphenes was shown to reduce collision rates on a navigation task in a simulated prosthetic vision study [111]. However, cueing is only provided upon request, with significant lag time, and relies on objects having high contrast with the environment.

2.4.3.2 Augmented Depth Obstacle Cueing

Augmented Depth is a depth-based representation designed for mobility scenarios, explicitly conveying the location of obstacles by boosting contrast with the ground for nearby obstacles [103]. Unlike previous work, obstacle detection is performed based on structure rather than appearance, which is more robust to variations in surface color and lighting conditions.

Concretely, obstacle detection on the input depth image D is implemented by

obtaining a ground plane mask G using iso-disparity contour analysis [101]. Under this scheme, a pixel is labelled as an obstacle if it is not contained in the ground plane. G is then used to form an augmented depth image D^* such that:

$$D^*(x,y) = \begin{cases} \gamma(\alpha x + \beta y + 1/D_0), & \text{if } G(x,y) \\ \lambda/D(x,y), & \text{otherwise} \end{cases} \tag{2.15}$$

where (α, β) defines the unit surface normal of the ground plane, D_0 is the depth of the point projecting along the planar surface normal, and $\gamma, \lambda > 0$ are visualization scale factors that control obstacle-ground contrast. Setting $\lambda > \gamma$ ensures that the contrast between obstacle pixels and ground pixels increases as object proximity increases. The vision processing method is defined by sampling the filtered augmented depth image at the projected phosphene locations:

$$B_{\text{augdepth}}(\varphi, D) = (D^* * L) \circ X(\varphi) \tag{2.16}$$

In a simulated prosthetic vision study [104], participants used augmented depth to complete a low-contrast obstacle avoidance task using a display consisting of 20-phosphenes arranged in a regular grid pattern. Use of this representation resulted in a significantly fewer collisions compared to intensity and raw depth, with mean per-trial collision rates of 0.545, 1.055, and 0.979 respectively. These results illustrate that structure-based obstacle detection and cueing can significantly improve functional outcomes in navigation tasks under low contrast conditions.

However, while augmented depth effectively displays obstacle locations in the scene, it does not convey the general structure of the scene. Knowledge of scene structure is useful for many tasks such as performing orientation or building a mental map of the environment. Furthermore, the the obstacle detection process requires surface fitting and assumes a single visible ground plane in the scene, restricting the range of application scenarios.

2.4.4 Limitations of SPV

Proposed vision representations are commonly evaluated through SPV studies. SPV aims to approximate an implant user's perception of electrode stimulation with an external rendering of phosphenes. SPV is a useful tool for estimating the effectiveness of vision processing methods without incurring the high overhead of performing a clinical trial, and can be used to provide a preliminary evaluation of experimental vision processing methods. However, SPV displays employ several simplifications that limit the accuracy of the approximation to the perceptual properties of a real implant display, and therefore results obtained using SPV should not be taken as an accurate reflection of implant user performance.

While the resolution and dynamic range of SPV displays can be set to match that of an implant, many other aspects of electrode stimulation perception are harder to model. Most often simulation models render phosphenes in a regular grid corresponding to electrode arrangements on implanted arrays, and phosphene appearance is approximated with abstract representations such as points [23], circles [39], or Gaussian spots [104]. However, the perceived locations of phosphenes is known to be irregular, and phosphene shape is non-uniform [140]. Furthermore, simultaneous electrode stimulations can entail interaction effects, stimulation tends to have a low refresh rate, phosphenes fade with repeated stimulation, and increased electrode resolution does not necessarily correlate with measured visual acuity [66, 165, 65]. While there is ongoing work to more accurately account for many of these conditions [54, 17], it is currently not possible to completely model the spatial and temporal properties of phosphenes resulting from electrode stimulation. Therefore, there are invariably effects during implant use that impact performance compared to observed SPV performance.

However, the relative performance of different visual representations can be expected to be similar between real implant use and SPV, since the approximations introduced by SPV would most likely affect different visual representations similarly. Therefore, SPV studies are generally comparative, providing a good idea of how dif-

ferent vision processing methods could be expected to perform relative to each other. For example, previous work has demonstrated that simulation results [90] can effectively predict relative performance between vision processing methods on a light localization task in subsequent patient trials [11].

2.5 Conclusion

This chapter has presented a review of structural salient object detection, structural edge detection, and vision processing for prosthetic vision. The literature review has found a number of limitations in existing work, which are as follows:

- Depth contrast, i.e. the distance between the foreground and background, forms the foundation of existing depth saliency methods. However, depth contrast methods ignore local surface shape, which is an important factor in making an object stand out from its surroundings.
- Furthermore, the saliency of an object should focus more on the structure of scene arrangements rather than than distances between scene components, as in depth contrast-based methods.
- Depth edge detection methods are either low-level or detect object contours; there is no focus on structurally salient edges such as a protruding corner from an object.
- There has been little application of high level structure-based computer vision techniques to form visual representations in prosthetic vision.
- No existing visual representations facilitate orientation and mobility with prosthetic vision while robustly conveying scene structure under low contrast conditions.

Chapter 3

Surface Irregularities

This chapter describes the surface irregularities method for detecting structurally salient regions in the scene. This method aims to address the thesis subproblem of identifying salient edges to convey scene shape, and has been developed for the task of orientation and mobility with prosthetic vision. Our approach is based on the idea that regions of irregular surface structure provide useful information about scene shape, while uniform and featureless regions of the scene are relatively undescriptive. An introduction to the problem is given in Section 3.1. The surface irregularities method is described in Section 3.2, and a quantitative assessment of boundary recall using surface irregularities is given in Section 3.3. The chapter is concluded in Section 3.4. Further evaluation of this method is available in Chapter 7, which presents a user study investigating the effectiveness of surface irregularities for navigation with SPV, and also discusses results from retinal implant users on a navigation task.

3.1 Introduction

The aim of this chapter is to develop a scene representation that conveys structurally irregular regions for navigation with prosthetic vision. Irregular surface regions contain information about the shape of the scene, and can be helpful for performing physical tasks such as navigation. For example, structural boundaries have irregular surface shape, and convey the general silhouette of a scene and the objects within it,

D	input depth image
w	window size
W	set of window sizes
$H_{w,x,y}$	histogram on window size w at image position (x, y)
r	neighbourhood radius
C_{SI}	cost metric between histograms of two windows
$\mathcal{S}_{\mathrm{SI}}$	single scale surface irregularities cost
S_{SI}	multi scale surface irregularities cost

Table 3.1. List of symbols used in this chapter.

which is useful for performing self-orientation or building a mental map of the surroundings in a prosthetic vision scenario. Also, knowledge of cluttered areas, which exhibit highly irregular surface shape, is helpful for planning safe routes when moving through the environment [102]. Conversely, smooth and uniform surfaces in the scene, such as the face of a wall or traversable space on the ground, provide less information about the overall structure of the scene. Therefore, navigation can be supported by highlighting irregular surface regions and ensuring that they are available within the capacity of the display.

Previous work in structural representations for prosthetic vision navigation has shown that artificially enhancing the contrast between obstacles and the ground plane can significantly improve the perception of small ground-based obstacles [104]. However, while this method conveys obstacle locations, it does not provide a more general depiction of scene structure, which is useful for interpreting the scene. Also, determination of the ground plane can be ambiguous, often requiring dominant surface assumptions which do not always hold when the camera is head-mounted, and scanning the scene.

In this chapter we propose a novel method for highlighting surface regions that have irregular structure compared to their surroundings. Unlike previous approaches, we achieve this without computing pixel-wise surface normals, estimating surface

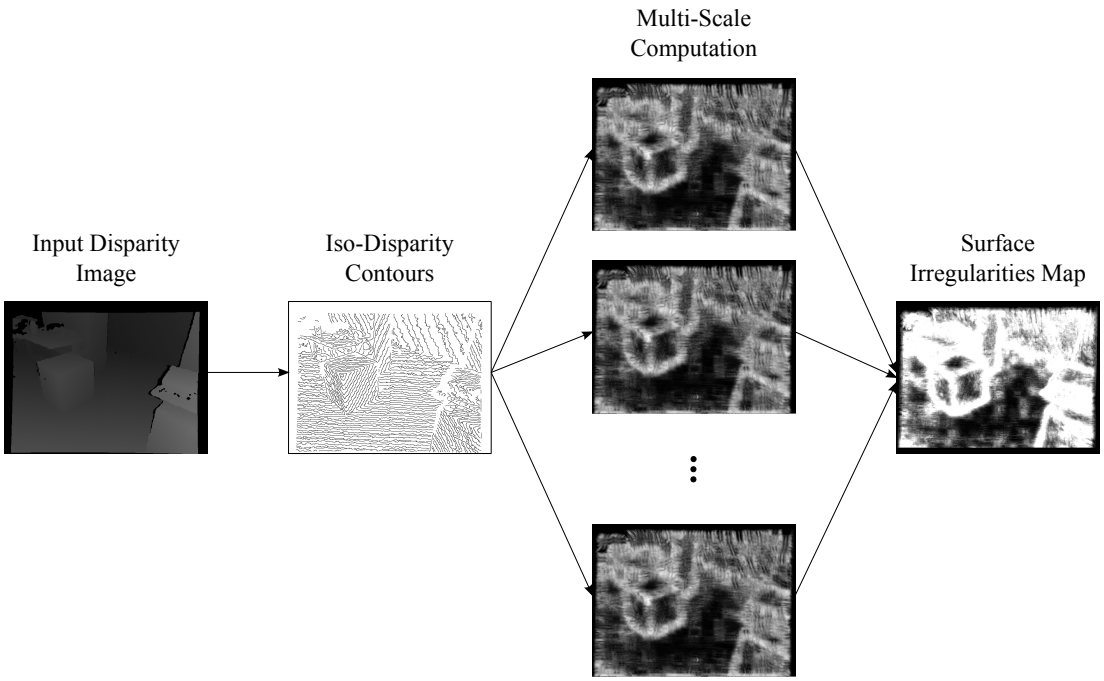


Figure 3-1. Overview of the pipeline of our method.

models, or use of appearance-based features from colour/intensity images. Rather, we exploit the arrangement of iso-disparity contours in depth images to statistically determine regions of structural significance in the scene, such as surface boundaries and general clutter. The use of iso-disparity contours has previously been reported for planar surface fitting [101]. Here, we do not explicitly model surfaces, but instead treat iso-disparity contour orientations as an observable feature in the depth input, from which smooth and non-smooth regions may be inferred. Our method is real-time, facilitating deployment as a scene representation for prosthetic vision devices. Results show that our method accurately and robustly highlights all obstructions in the scene, as well as major surface boundaries. Qualitative examination of the resulting visual representation in simulated prosthetic vision demonstrates the potential of our approach to support safe mobility with current and near-term visual prostheses.

3.2 Surface Irregularities Detection

Given as input a dense, discretised depth map D, we compute a surface irregularities map S_{SI}, which reflects how structurally significant the pixel is relative to the whole scene. Pixels corresponding to clutter or surface boundaries are expected to have a higher surface irregularities score than pixels on smooth surfaces. This score is calculated in four key steps:

1. iso-disparity contour extraction and multi-scale histogram computation,
2. fixed-scale surface irregularities calculation,
3. multi-scale surface irregularities fusion, and
4. contour-disparity ratio and gradient magnitude adjustment.

Each of these steps is outlined below.

3.2.1 Extraction and Multi-scale Histogramming of Iso-disparity Contours

The first step is the extraction of iso-disparity contours which form the basis of the surface irregularity computation. Canny edge detection [22] is applied to D to produce a binary image of iso-disparity contours, determined from the boundary between discrete depth levels (see Figure 3-1). These iso-disparity contours are then divided into linear piecewise segments, in order to estimate the local orientation of each contour point. This is achieved by iteratively forming straight line segments on contour points until an error of 4 pixels is exceeded, at which point the segment is stored and the process repeated.

A multi-scale sliding window is passed over the iso-disparity image to determine the local distribution of iso-disparity orientations at each position. Orientations within each window are counted into one of 9 histogram bins. This number of bins was found to provide a good balance between descriptiveness and robustness to noise.

For a window with side length w at position (x, y), the resulting histogram is denoted as $H_{w,x,y} : [1...B] \to \mathbb{R}$.

3.2.2 Window Surface Irregularities Computation

The difference in surface structure between two regions is computed by comparing their iso-disparity contour orientation histograms. We define a smoothed cost metric between two windows of scale w positioned at (x, y) and (u, v) as

$$C_{\mathrm{SI}}(w, x, y, u, v) = \log\left(1 + \frac{\sum_{b=1}^{B} |H_{w,x,y}(b) - H_{w,u,v}(b)|}{\sum_{b=1}^{B} \max\left(H_{w,x,y}(b), H_{w,u,v}(b)\right)}\right) \tag{3.1}$$

This cost reflects the non-overlapping portion of the histograms as a ratio to total size. A nonlinear function is used to balance and reduce the effect of large costs.

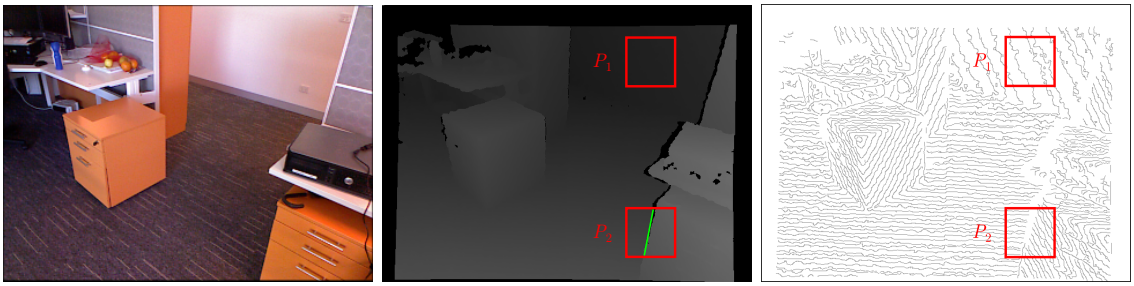
The surface irregularities score of a given window is taken as the minimum cost between the window and neighbouring windows in the same scale. Thus, given a scale w, window position (x, y), and window neighbourhood radius r, the single-scale score of the window is given by

$$\mathcal{S}_{\mathrm{SI}}(w, x, y, r) = \min_{-r \le i \le r, -r \le j \le r} C_{\mathrm{SI}}\left(w, x, y, x + i \cdot w/2, y + j \cdot w/2\right) \tag{3.2}$$

3.2.3 Multi-scale Surface Irregularities Fusion

Surface irregularities scores are aggregated across multiple window sizes in order to provide a degree of robustness to object scale and image noise. The multi-scale surface irregularities score of a given point is computed by first finding the window score of the point for a number of different scales, and then merging the results via a weighted average by window occupancy. Thus, defining W as the set of window sizes and $C(w, x, y)$ as the number of iso-disparity contour pixels contained in the window of size w centred on (x, y), the surface irregularities score for a pixel is given by

$$S_{\mathrm{SI}}(W, x, y, r) = \sum_{w \in W} \frac{C(w, x, y)}{w^2} \mathcal{S}_{\mathrm{SI}}(w, x, y, r) \tag{3.3}$$



(a) RGB (b) Depth (c) Iso-disparity Contours

Figure 3-2. Example scene with two windows P_1 and P_2, showing areas in the scene where contour-disparity ratio and gradient magnitude adjustment would be applied. P_1 contains a fronto-parallel surface with a low iso-disparity contour to window area ratio. P_2 contains a depth discontinuity edge with high depth gradient shown in green.

Setting $r = 10$ and $W = \{20, 30, 40\}$ was found to provide strong detection of relevant surface irregularities for general use.

3.2.4 Contour-disparity Ratio and Gradient Magnitude Adjustment

We perform two post-processing steps to improve the multi-scale surface irregularities image: lowering the response on fronto-parallel surfaces, and increasing the score at depth discontinuity edges.

Fronto parallel surfaces pose a challenge since their iso-disparity contours are relatively sparse. We detect such surfaces explicitly by computing the local ratio of iso-disparity pixels to the total number of valid depth values in a window. If the ratio is near-zero, then we assume the region represents a near-frontal surface. See P_1 in Figure 3-2 for an example of a fronto-parallel surface.

Gradient magnitude thresholding of D was performed to explicitly identify depth discontinuity edges in order to increase surface boundary recall. Any pixel in the normalised gradient magnitude image of D with a value above a small threshold of 0.0005 was assumed to represent part of a depth discontinuity edge. See P_2 in Figure 3-2b for an example of a depth discontinuity edge.

3.3 Experiments and Results

We now provide a validation of the proposed surface irregularities method for prosthetic vision navigation. First, the run time of the surface irregularities method for the experiments is presented. After this, quantitative evaluation is performed to measure how well the method is able to detect boundaries in the scene and how well it identifies traversable space compared to the state-of-the-art Augmented Depth structural scene representation. Finally a qualitative examination of our method is presented.

3.3.1 Run Time

An implementation of our surface irregularity method takes less than 60ms per frame on an Intel i3 2.9 GHz processor. This allows the method to be used to perform real-time vision processing in prosthetic vision scenarios.

3.3.2 Quantitative Comparison: Surface Boundary Recall

The output of the proposed method is inherently qualitative. However, to provide a preliminary validation of the appropriateness of our approach for the needs of mobility, we perform two quantitative measures. These measures are calculated according to ground truth obtained via hand-labelling of pixels belonging to surface boundaries, and pixels belonging to the ground plane.

- Surface boundary recall rate (SBRR): the proportion of correctly labelled pixels along surface boundaries for the surface irregularities map:

$$\text{SBRR} = \frac{TP_{\text{boundary}}}{TP_{\text{boundary}} + FN_{\text{boundary}}}, \tag{3.4}$$

  where TP_{boundary} and FN_{boundary} are the number of true positives and false negatives respectively, computed from the surface boundary ground truth. This metric measures the ability of the method to identify all surface irregularities in the scene. A higher SBRR score implies that the method is less likely to miss critical structure such as small trip hazards.

	Perturbance $t = 0.4$		Perturbance $t = 0.7$		Plane Fitting	
Image Id	SBRR	GPMR	SBRR	GPMR	SBRR	GPMR
(I)	0.98	0.18	0.62	0.04	0.74	0.00
(II)	0.94	0.21	0.54	0.05	0.52	0.05
(III)	0.90	0.25	0.52	0.07	0.71	0.03

Table 3.2. Quantitative results showing surface boundary recall rate (SBRR), and ground plane mislabel rate (GPMR) for the perturbance map (thresholds 0.4 and 0.7), and plane fitting.

- Ground plane mislabel rate (GPMR): the proportion of ground plane pixels incorrectly labelled as surface boundaries/clutter:

$$\text{GPMR} = \frac{TP_{\text{plane}}}{TP_{\text{plane}} + FN_{\text{plane}}}, \tag{3.5}$$

  where TP_{plane} and FN_{plane} denote the number of true positive and false negative detections of ground plane pixels. This score reflects the capacity of the method to distinguish traversable space from boundaries and clutter.

The SBRR and GPMR metrics are calculated using a thresholded binary segmentation of the normalised surface irregularities image by a threshold t. Two thresholds are used in the evaluation, $t = 0.4$ and $t = 0.7$, to measure performance at relatively low and high precisions respectively.

Table 3.2 shows results for a small set of test images acquired using a Microsoft Kinect sensor. Here we report SBRR and GPMR results using the proposed surface irregularities map. As a point of comparison, we also include results obtained from the plane-fitting technique described in [101]. While not a perfect benchmark of our approach, we include this to provide some indication of how the two approaches compare for distinguishing traversable and non-traversable space in the scene. Figure 3-3 shows the test images and resulting binary thresholded surface irregularities map ($t = 0.4$) for each.

Most notably, the $t = 0.4$ surface irregularities map achieves an SBRR above 90% for all images. GPMR results for the surface irregularities map are less impressive for $t = 0.4$, but improve significantly for $t = 0.7$, indicating a clear trade-off between

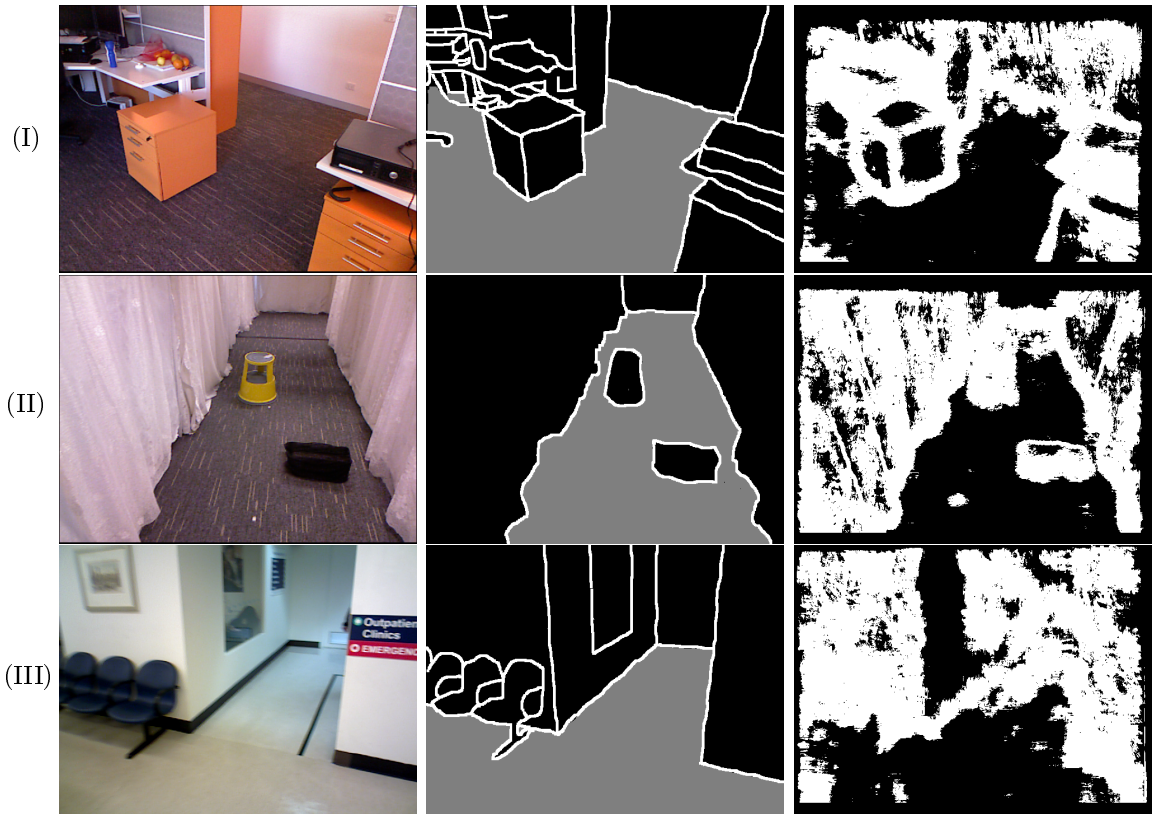
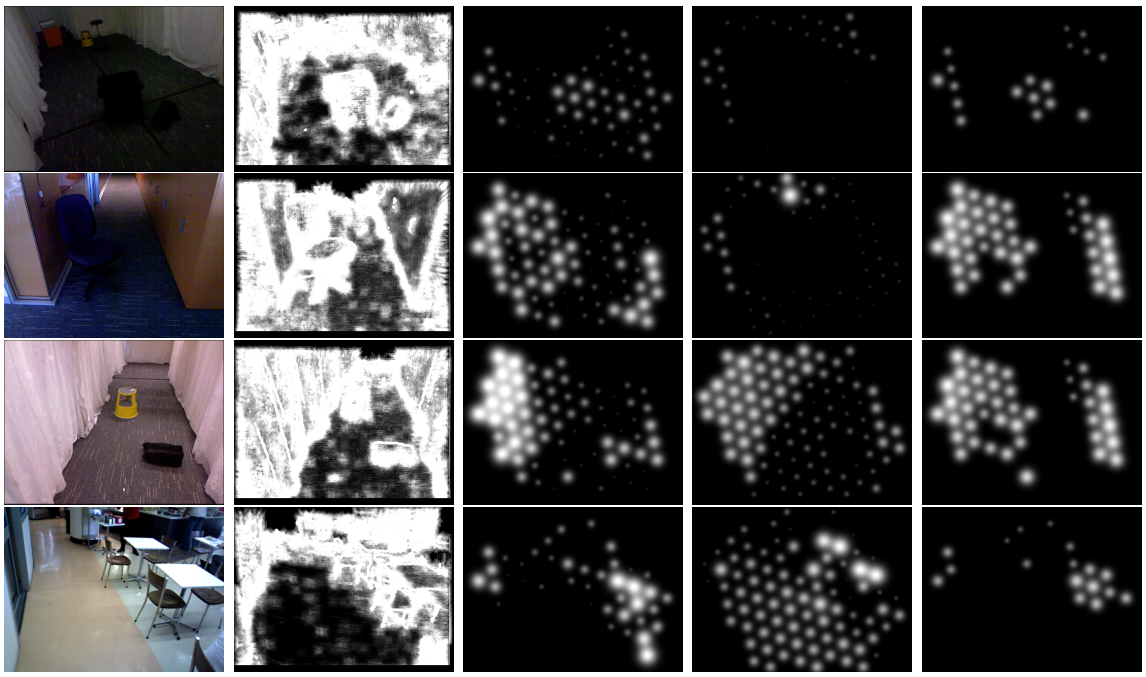


Figure 3-3. Images used for quantitative results in Table 3.2: (a) RGB image, (b) ground truth, in which boundary pixels are labelled white and ground plane pixels are labelled gray, and (c) $t = 0.4$ binary thresholded surface irregularities map.

recall rate and mislabelling. Visual inspection of the 0.4 surface irregularities segmentation shows that in all images, mislabelling is primarily due to the thickness of boundary segmentations. Away from the ground surface boundaries, mislabelling is rare. The comparatively lower SBRR results for plane fitting are unsurprising given the method makes no explicit attempt to detect boundaries. Thus, ground plane labels can easily bleed across boundaries. Overall, quantitative results suggest the proposed surface irregularities map provides comparable differentiation of ground surface and obstacles to plane fitting, but with significantly higher boundary recall rates. These results suggest the proposed method is well-suited to the needs of safe mobility with prosthetic vision.



(a) (b) (c) (d) (e)

Figure 3-4. Qualitative results: comparison of surface irregularities with state-of-the-art prosthetic vision scene representations on a simulation of a 98-phosphene display. (a) RGB image (b) surface irregularities map (c) surface irregularities SPV (d) standard intensity SPV (e) Augmented Depth [103] SPV. SPV denotes simulated prosthetic vision.

3.3.3 Qualitative Comparison

This section provides a qualitative comparison of our surface irregularities method with the standard downsampled intensity and Augmented Depth visual representations when rendered using SPV. The surface irregularities visual representation is defined by the following brightness function:

$$B_{\mathrm{SI}}(\varphi, D) = ((S_{\mathrm{SI}}(D)./D) * L) \circ X(\varphi), \tag{3.6}$$

where φ is the output phosphene, D is the input depth image, $S_{\mathrm{SI}}(D)./D$ is the surface irregularities map with pointwise scaling by inverse depth, L is the Lanczos2 kernel, $X(\varphi)$ is the projected location of φ in image space. Thus, the brightness of each phosphene is equal to the filtered surface irregularities map scaled by inverse depth for visualisation. The standard downsampled intensity representation directly

conveys scene intensity at the projected phosphene locations, and the Augmented Depth representation conveys scene disparity while suppressing phosphenes on the ground plane, increasing the contrast of objects resting on the ground. The brightness functions for the standard intensity and Augmented Depth visual representations are given in Equations 2.14 and 2.16 respectively.

Figures 3-4a and 3-4b show sample images and the resulting surface irregularities map obtained using the proposed method. It can be seen that the surface irregularities map provides clear delineation between clutter in the scene and the dominant smooth surfaces. In particular, small ground obstacles such as the dark box in row 3 produce a high response relative to the ground plane. Non-ground smooth surfaces such as the walls in row 2, the seat cushion in row 2, and table tops in row 4 are all assigned low surface irregularities scores relative to their boundaries, and other clutter in the scene.

Figures 3-4c, 3-4d, and 3-4e display the associated visual representations rendered with simulated prosthetic vision. From observation it can be seen that both the surface irregularities-based representation, and Augmented Depth, provide reasonable distinction between the ground and obstructed space. Notably, however, the surface irregularities based representation captures more of the structural detail such as that present in rows 2 and 4. It also highlights the black obstacle in the foreground of the scene in row 3, which is missed by both Intensity and Augmented Depth. As expected, low-contrast objects are generally not visible in the intensity-based representation.

3.3.4 Discussion

These results demonstrate the effectiveness of the proposed surface irregularities map for detecting and emphasising structurally significant regions in the scene. We have shown that regions of irregular surface shape better identify important surface boundaries than existing structural scene representations, while also preserving information about traversable space. This provides insight into the thesis subproblem of identifying the types of structural features that are important for understanding a scene. While plane-fitting methods can generally be expected to achieve greater ground pixel

labelling than our approach, the surface irregularities map provides a significantly better recall rate for surface boundaries. This is arguably the more relevant performance indicator for safe mobility with prosthetic vision, ensuring all boundaries present in the input scene are preserved, and no potential trip hazards are missed. It is important to note that the surface irregularities approach makes no limiting assumptions about the scene, such as planarity or colour contrast; it simply characterises smoothness. The surface irregularities map also provides a richer description of the scene, characterising all regions of clutter, and conveying structural boundaries as well as object shape.

3.4 Chapter Summary

In this chapter we have presented a novel approach for detecting salient edges for the purpose of facilitating safe mobility with prosthetic vision. Ours is the first method that conveys structural saliency directly to the user, enhancing the display of not just obstacle locations but also general scene structure, without making limiting assumptions about the scene. Our method identifies the types of structural features are important for understanding a scene, with regions of irregular surface structure conveying scene shape through surface boundaries and traversable space.

We have proposed a method of measuring structural saliency using iso-disparity contours. Regions of significant structural change are detected via a cost function based on local comparisons of iso-disparity contour orientations. Through this, structurally interesting features such as surface boundaries and general clutter are extracted and emphasised in the output visual representation.

Our approach is real-time, avoids assumptions of high contrast environments, and removes the need to explicitly reconstruct the scene via surface fitting. Our results demonstrate how the proposed Surface Irregularities map may be used to emphasise all surface boundaries and clutter in the scene, robustly and efficiently. More generally, the proposed approach demonstrates how analysis of scene structure using depth data may provide advantages for supporting mobility with near-term

prosthetic vision devices. Further evaluation of the surface irregularities method through a prosthetic vision user study is conducted in Chapter 7.

Chapter 4

Deep Structural Edges

This chapter continues the investigation of the salient edge thesis subproblem, and describes a new salient edge detection system that was developed to improve the surface irregularities method from Chapter 3. Detection of surface irregularities such as edges mimics low level biological processes that support perception of the environment [130]. Biological vision, however, also incorporates high level information from task knowledge and memory that influences the importance of perceived edges. This leads to certain structurally salient edges, for example the boundaries of a room or dangerous protruding edges, being more likely to contribute to high level decisions when performing physical tasks such as navigation. In this chapter, we aim to extract structurally salient edges from the environment using learned high level information, in order to enable our prosthetic vision scene representation to more closely emulate the mechanisms of the human visual system.

Specifically, a fully convolutional neural network is used to model salient edges from a proposed minimal depth input encoding called the Depth Surface Descriptor (DSD). Section 4.1 outlines the problem and our approach. In Section 4.2, we introduce the DSD feature, and Section 4.3 provides a comparison of the raw DSD feature with the current state-of-the-art depth input encoding. Our CNN edge detection system is introduced in Section 4.4, Section 4.5 describes the experiments and the implementation details, and Section 4.6 presents the experimental results of our edge detection system. Evaluation is performed on a standard depth contour

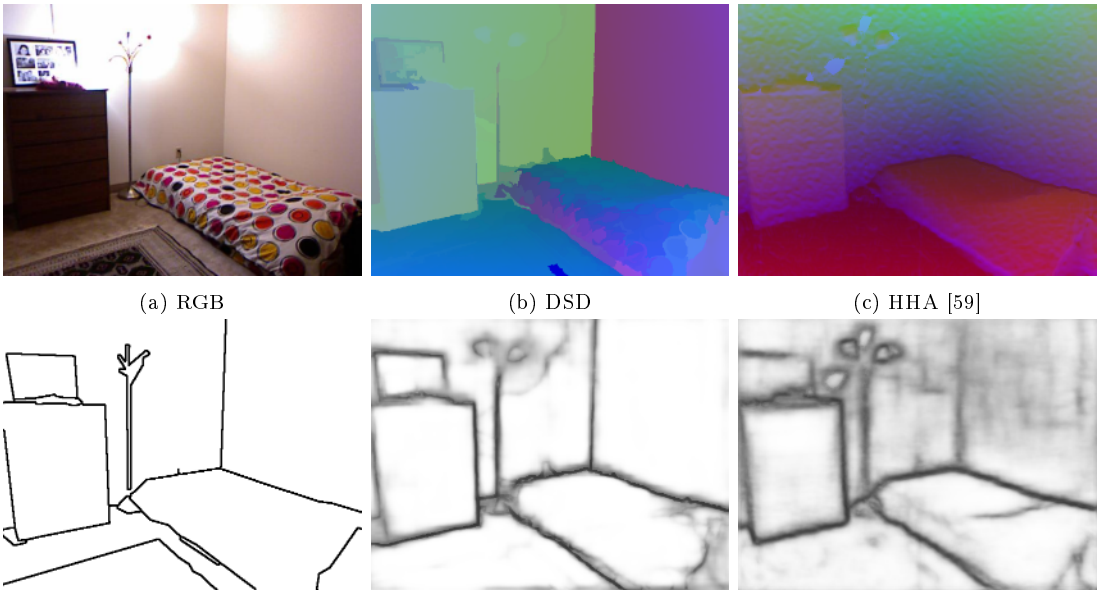
D	input depth image
X	input 3D surface
S^2	unit sphere
$(\mathbf{x}_0, \mathbf{y}_0, \mathbf{z}_0)$	camera coordinate frame
N	Gauss map
$\mathbf{u}, \mathbf{v}$	frame of reference for Gauss map approximation
$\mathcal{N}$	approximation of Gauss map
$\tilde{N}$	discretised normal map from sensor data
$\mathcal{I}$	ideal input encoding for edge detection
σ_L, σ_U	Canny thresholds

Table 4.1. List of symbols used in this chapter.

detection dataset as well as a custom prosthetic vision navigation dataset, which was constructed to enable salient edge detection assessment for assistive navigation scenarios. Section 4.7 concludes the chapter.

4.1 Introduction

Knowledge of scene structure is important for mobility with retinal implants where the bandwidth of image information that can be represented per frame is highly restricted, and displaying only salient structure allows the user to interpret the scene around them [104]. While low level structural edge detection methods such as surface irregularities from Chapter 3 accomplish this to some extent, these methods generally detect all irregular structure or edges without assessing their relevance in a prosthetic vision navigation scenario. For example, the locally strong structural texture on the curtains in Figure 4-6 elicits a high response from low level edge detectors, and can increase the difficulty of interpreting the scene on a prosthetic vision display. Results from low-level detection methods can be improved by including scale, context, and other high level information to suppress structural edges that are not significant



(a) RGB (b) DSD (c) HHA [59]

(d) Ground Truth (e) Ours (DSD) (f) HED-HHA[157]

Figure 4-1. HED [157] output when trained on our DSD feature compared to the standard HHA [60] depth feature. Note that our feature provides a better representation of the scene structure, e.g. the corner between the two walls at the right of the image, and is significantly less affected by sensor noise, allowing the CNN to better model edge structure and thus produce a more accurate edge map.

for interpreting the scene. In this chapter, we revisit finding structural edges that are significant for 3D scene understanding and mobility by learning this high level information to improve detection of salient edges.

Recent contour detection methods [157, 107] incorporate high level information to improve edge detection performance, however these methods suppress structural edges that are internal to objects. Edges such as the corner between two walls could be regarded as internal to the wall object, but are crucial to indoor scene understanding [129], and provide important cues for assistive navigation. Similarly, a large object with a leading edge may protrude substantially from the object, posing a collision risk.

State-of-the-art structural edge detectors all operate based on the HHA depth encoding [157, 16, 131]. While this encoding is useful for contour detection, it is less suited to structural edge detection since it does not incorporate a full model of curvature. For example, the HHA feature does not directly represent vertical joins

between two surfaces, such as the boundary between adjacent walls, or the corner on a wardrobe (see Figure 6-2). These types of edges are a common occurrence in indoor scenes and are usually salient.

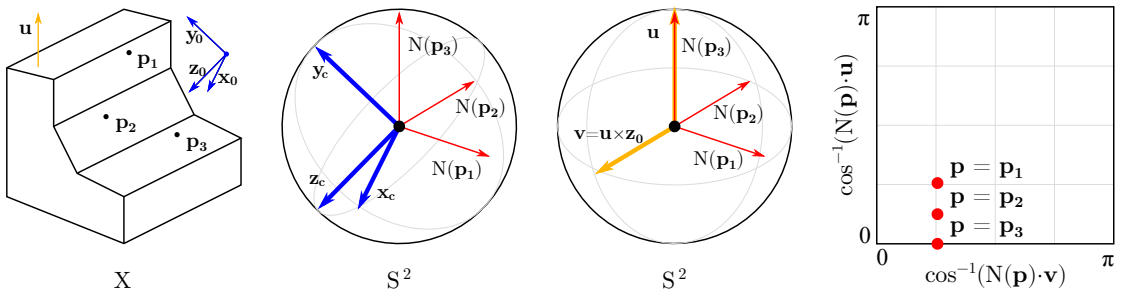
Hence, this chapter investigates improving the recovery of structural edges that are significant for 3D scene understanding and mobility from RGB-D input. In particular, we contribute a new depth input encoding that is suited to finding structural edges in RGB-D images that relate to the important aspects of scene structure. We incorporate this depth encoding into an end-to-end fully convolutional neural network framework, extracting salient edge structure from the scene using learned high level information.

4.2 DSD Feature

In this section we introduce our proposed depth feature, the Depth Surface Descriptor (DSD), which aims to provide a minimal encoding of depth input that captures the distinguishing surface geometry of structural edges, and suppresses sensor noise and other non-edge structure.

We are interested in depth edges as opposed to appearance edges that are treated separately in our architecture. Depth edges arise for only two reasons, a first order discontinuity in the surface (*i.e.* a crease edge), or a depth discontinuity in the surface (*i.e.* a step edge). To develop a structural edge detector, we require that it can identify these phenomena regardless of the nature of the appearance or embedding of the surface.

Classically, Gaussian curvature encodes the intrinsic curvature of a surface regardless of embedding [52]. Hence, two principal curvatures are all that are required to encode a surface. For a single view input scene $X \subset \mathbb{R}^3$, this information can be represented as a Gauss map $N : X \to S^2 = \{\mathbf{x} \in \mathbb{R}^3 \mid \|\mathbf{x}\| = 1\}$ defined on the camera coordinate system, which maps a surface point $\mathbf{p} \in X$ to the point $\mathbf{n} \in S^2$ on the unit sphere corresponding to the surface normal at $\mathbf{p}$. Note that X can be obtained from a depth image D by back-projecting each pixel in D to the 3D surface point that it represents.



(a) (b) (c) (d)

Figure 4-2. (a) Input surface X with three marked points $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$. The camera coordinate frame $(\mathbf{x}_0, \mathbf{y}_0, \mathbf{z}_0)$ and gravity direction $\mathbf{u}$ are shown. (b) Gauss map N for each point shown with the camera coordinate frame. (c) Gauss map N for each point shown with the reference frame vectors $\mathbf{u}$ and $\mathbf{v}$. (d) The Gauss map approximation $\mathcal{N}$ of each point computed from the reference frame vectors $\mathbf{u}$ and $\mathbf{v}$.

Since we seek a minimal encoding, we use an approximation function $\mathcal{N}$ defined as follows:

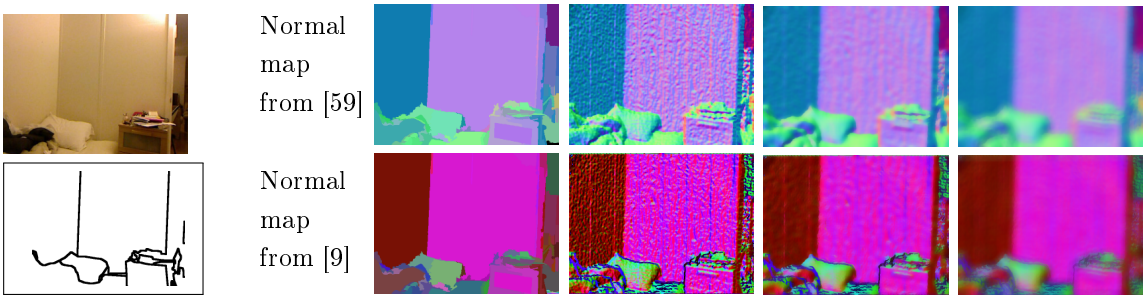
$$\mathcal{N}(N(\mathbf{p})) = \left(\cos^{-1}\left(N(\mathbf{p}) \cdot \mathbf{u}\right),\ \cos^{-1}\left(N(\mathbf{p}) \cdot \mathbf{v}\right)\right), \tag{4.1}$$

where $\mathbf{u}, \mathbf{v} \in S^2$ are fixed and orthogonal. $\mathcal{N}$ is injective, i.e. $\mathcal{N}(\mathbf{a}) = \mathcal{N}(\mathbf{b}) \rightarrow \mathbf{a} = \mathbf{b}$. Too see this, consider that

$$\cos^{-1}(\mathbf{a} \cdot \mathbf{u}) = \cos^{-1}(\mathbf{b} \cdot \mathbf{u}) \rightarrow \mathbf{a} \cdot \mathbf{u} = \mathbf{b} \cdot \mathbf{u} \rightarrow \mathbf{u} \cdot (\mathbf{a} - \mathbf{b}) = 0, \tag{4.2}$$

since we are only interested in the range $[0, \pi]$ where $\cos^{-1}$ is bijective. This implies that $\mathbf{u}$ is perpendicular to $\mathbf{a} - \mathbf{b}$, i.e. $\mathbf{a}$ and $\mathbf{b}$ both lie on a circular path on the unit sphere around $\mathbf{u}$. Similarly, $\mathbf{a}$ and $\mathbf{b}$ both lie on a circular path around $\mathbf{v}$. Since $\mathbf{u} \neq \mathbf{v}$ and $\mathbf{u} \neq -\mathbf{v}$, the intersection of these circular paths have two solutions at antipodal points on the unit sphere, giving either $\mathbf{a} = \mathbf{b}$ or $\mathbf{a} = -\mathbf{b}$. Since all normals in the depth image must have a positive dot product with the viewing ray, it follows that $\mathbf{a} = \mathbf{b}$. Therefore $\mathcal{N}$ does not reduce the discriminability of the representation.

While $\mathcal{N}$ is sufficient for detection of structural edges intrinsic to a surface, it does not naturally capture step edges. Step edges are viewpoint-dependent discontinuities in scene depth induced by occlusion in single view representations, and do not result in



(a) RGB + GT (b) $\tilde{\mathcal{N}}$ (c) Unfiltered ($\tilde{N}$) (d) $\sigma = (20, 200)$ (e) $\sigma = (40, 400)$

Figure 4-3. Visualisation of of our surface patch mapping function $\tilde{\mathcal{N}}$ and aggressive bilateral smoothing of pointwise normals computed using two different methods [59, 9]. Our method mitigates noise within surface patches while maintaining contrast between regions bordering structural edges.

a visible change of surface normal (e.g., stairs viewed from directly above). Therefore, we must include the original depth map D in our minimal encoding in order to represent these types of edges.

Our minimum encoding thus consists of absolute depth and surface normals. Next we present how we compute stable surface normals.

4.2.1 Region-Based Normal Smoothing

Sensor noise has an adverse effect on the image representation, often introducing false positives during edge detection. The effects of sensor noise are magnified in surface normal estimations from depth sensors, since surface orientation is a first order property of the sensor output. As shown in Figure 4-3, a seemingly flat surface can have a wide distribution of surface orientations due to a small amount of noise in the depth reading. Thus the noisy discretised normal map $\tilde{N}$ computed from sensor data is a poor approximation to N and does not accurately express the surface structure of the scene.

Low level image filtering [59, 9, 53] can address this issue to an extent by smoothing spurious local normal variations, but still leaves a considerable amount of noise in the input. Furthermore, over filtering will blur the structural boundaries of the scene, reducing edge localization accuracy, as shown in Figure 4-3. Due to the unknown required scale of surface curvature, filter size cannot be defined *a priori*.

We reduce the effect of sensor noise by performing region-based smoothing of the point-wise normal image. First, we over-segment the image into surface patches using the Mean Shift algorithm [36]. Following this, the surface normal of each point within a patch is replaced with the average surface normal of the patch. This maps regions with consistent surface orientation to a single representative normal value, smoothing normals within a surface while maintaining contrast between surfaces that border structural edges.

4.2.2 Normal Computation Frame of Reference

The ground orientation is a key piece of semantic information in many scenes, as it provides an absolute reference point for object surfaces in the scene. For example, a boundary between the ground and a vertical surface, or between two walls, may be more likely to be labelled as salient, particularly for tasks such as mobility. We parameterise $\mathcal{N}$ with respect to the ground plane by fixing the first coordinate axis $\mathbf{u}$ to the inferred direction of gravity. To provide a stable reference frame, the second axis $\mathbf{v}$ is set to be orthogonal to both the camera axis $\mathbf{z}_0$, *i.e.* the direction the camera is facing, and $\mathbf{u}$.

$$\mathbf{v} = \mathbf{u} \times \mathbf{z}_0. \tag{4.3}$$

This increases the amount of information encoded in our minimal representation with no representation cost, providing a stable reference frame from which further relationships between edges and scene structure can be inferred.

4.3 Comparison of Raw DSD and HHA Features

Before presenting our edge detection system, we first compare the suitability of the proposed DSD encoding with the state-of-the-art HHA encoding for edge detection. Specifically, this section presents a quantitative and qualitative comparison of the DSD and HHA depth encodings.

4.3.1 Quantitative Comparison

We now present a quantitative evaluation of the suitability of the DSD and HHA features for structural edge detection. The first step is to determine a metric for how appropriate the encoding is for edge detection. We note that the suitability of a representation for edge detection is reflected by how clearly boundaries are distinguished in the input encoding.

With this in mind, consider two adjacent regions P, Q in some scene represented by an *ideal* input encoding $\mathcal{I}$ for edge detection. If a boundary exists between P and Q then the two regions should be clearly separable in feature space, i.e. $|\mathcal{I}(P) - \mathcal{I}(Q)| = 1$. Conversely, if the two regions are identical in feature space $|\mathcal{I}(P) - \mathcal{I}(Q)| = 0$ then there is no boundary between the regions.

In this situation, applying a naive edge detector such as Canny [22] to the ideal encoding would yield a perfectly accurate edge map. Thus, in this section we measure the effectiveness of different depth encodings at representing edges by evaluating the edge maps produced when applying Canny to the encoded image.

The Canny parameters for different depth encodings are set using a grid search. The lower σ_L and upper σ_U Canny thresholds are selected from $\sigma_L \in \{20, 40, 60\}$ and $\sigma_U \in \{50, 100, 150\}$ to give the best performance.

4.3.1.1 Evaluation Metrics

We use three standard performance metrics for edge detection evaluation. These are the F-score for the best threshold over the dataset (ODS), best per image threshold (OIS), and average precision (AP).

4.3.1.2 Experiments and Results

We apply the Canny edge detector to the images on SUNRGBD, encoded with our DSD feature and the HHA feature, as in [94].

As shown in Table 4.2, the DSD feature provides a better representation of scene geometry for structural edge detection. This in particular supports learning a bet-

Method	ODS	OIS	AP
HHA	.422	.427	.012
DSD	**.489**	**.493**	**.025**

Table 4.2. Raw encoding results on SUNRGBD.

ter structural edge model, as demonstrated by our improved results in Section 4.6. Furthermore, this implies that generally the DSD feature is preferable to HHA when used in systems that require structural edge detection.

4.3.2 Qualitative Comparison of DSD and HHA

Figure 4-4 shows some DSD and HHA encodings for a number of different scenes from SUNRGBD. Also shown are the edge maps from the Canny edge detector applied to these maps. Note that surfaces bordering structurally significant boundaries generally have greater contrast in the DSD encoding compared to the HHA encoding. For example, in the first row of Figure 4-4, the wall edges are clearly represented in the DSD encoding, while they are not directly represented in the HHA encoding. This is reflected in the corresponding Canny edge maps, demonstrating that our feature is able to capture a wider range of structure than HHA.

The DSD feature also reduces the effect of depth sensor noise. An example of this can be seen in rows 3 and 4 of Figure 4-4. Here the HHA feature exhibits a high amount of noise, as seen on the side of the cabinet closest to the camera in row 3, and on the chairs in row 4. Noise is also usually present on distant surfaces in the HHA feature. In the DSD feature surface normal noise is largely eliminated via region-based normal smoothing. This produces a cleaner feature map, as seen in the generated Canny edge maps of row 3 and 4.

From Figure 4-4 we can see that in general surfaces bordering structural boundaries are not separable from each other in the HHA feature map. Even with the best Canny parameters, the difference between different surfaces is often a similar magnitude to noise, and thus there is no fixed threshold that can effectively distinguish surface boundaries to noise. Our DSD feature, on the other hand, is designed

to be a minimal encoding that captures the full surface geometry profile, and thus the structure bordering significant boundaries is generally separable. As a result, the DSD feature produces better final accuracy than HHA in the learned model of the main paper.

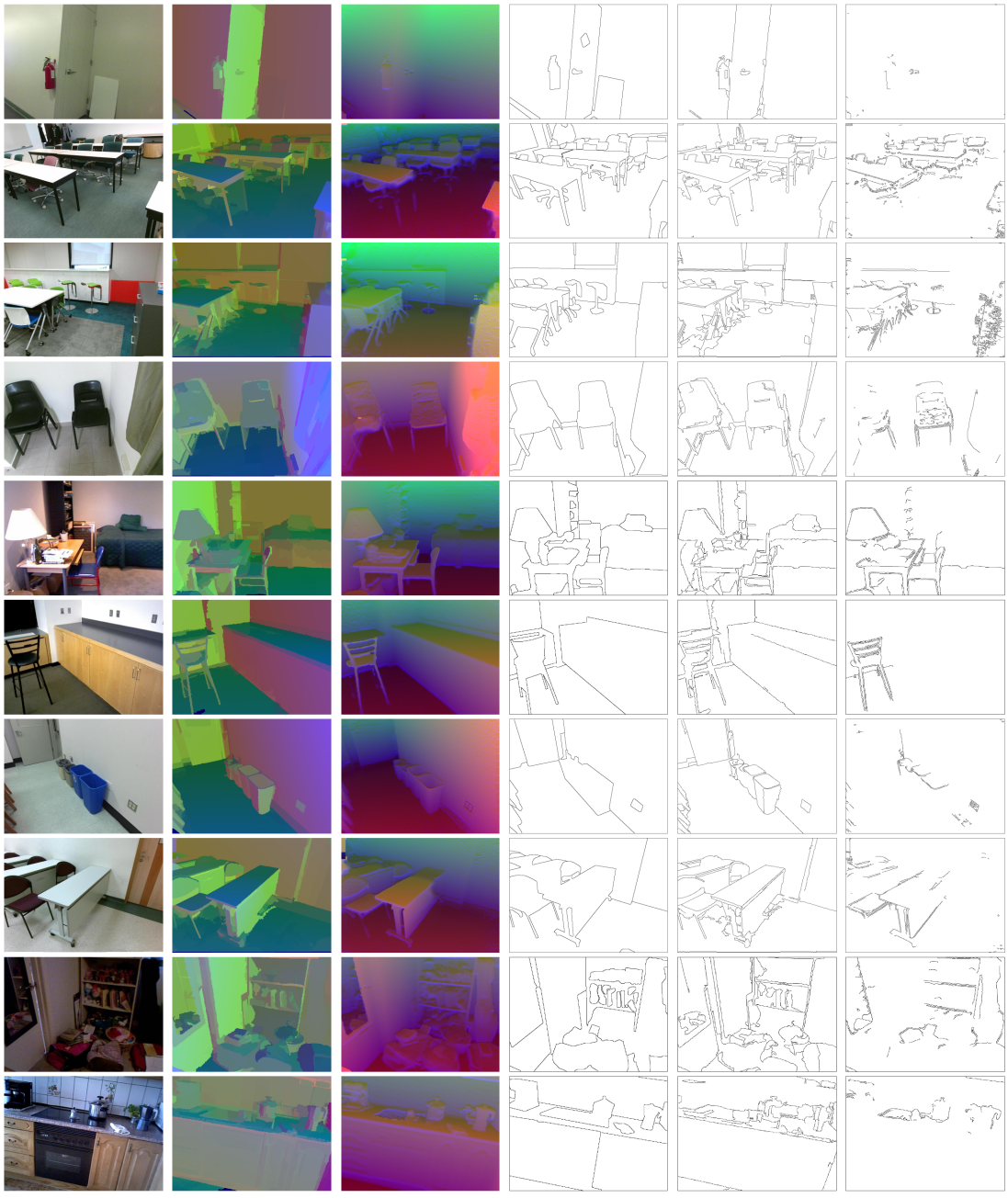
4.4 Edge Detection System

This section describes our method of detecting structural edges from the scene. To extract edges from the scene, we require a spatial operator to perform detection on the minimal depth encoding. The desired operator must account for all surface shapes, such as two corrugated iron fences that abut at an angle, or a corner in rippled curtains (see Figure 4-6). In addition, sensor noise is complex and scale is problematic, in short, "mathematics has nothing to say about scale" - O. Faugeras [47]. A rippling curtain does have changing curvature, but it is the joint between the surfaces that would be considered structurally salient by humans for most tasks (see Figure 4-6). For these reasons, low level edge operators such as surface irregularities are generally unsuitable for the task.

Hence we take the approach of applying a deep CNN that incorporates high level semantic and context information into account as the spatial operator to detect structurally salient edges. An advantage of deep CNNs for such problems is that the encoding weighs depth values from the entire image and so supports a multi-scale framework. Further, contour processing generally employs a broader region of support to suppress noise as well as a local gradient operator to find the edge.

4.4.1 Network Architecture

We use the VGG-16 fully convolutional network as the base architecture for testing the DSD encoding. Since our main contribution is the DSD encoding, the selection of VGG-16 provides fair comparison of this encoding with existing methods. We trim the fully connected layers of VGG and incorporate deep supervision by adding a side output to the last convolutional layer of each of the five VGG blocks, as in [157].



(a) RGB (b) DSD (c) HHA [60] (d) GroundTruth (e) Ours (DSD) (f) HHA [60]

Figure 4-4. Edge maps obtained by applying the Canny filter applied to the DSD and HHA [60] depth encodings, with the best Canny parameters found via grid search. This gives a measure of how effectively each feature encodes geometric surface information for edge detection. Note that in the DSD encoding, surfaces bordering structurally significant boundaries generally have greater contrast.

We will now give a brief overview of the objective function of the network. For more details, please see [157].

Let $\mathbf{W}$ denote the collection of standard network parameters. Suppose we have M side output layers, where each side output is associated with a classifier with corresponding weights $\mathbf{w} = (\mathbf{w}^{(1)}, ..., \mathbf{w}^{(M)})$. For a given input $X = \{x_j, j = 1, ..., |X|\}$ and ground truth $Y = \{y_j, j = 1, ..., |X|\}$, the image-level loss function for the side outputs is given by:

$$\mathcal{L}_{\text{side}}\left(\mathbf{W}, \mathbf{w}\right) = \sum_{m=1}^{M} \alpha_m l_{\text{side}}^{(m)}\left(\mathbf{W}, \mathbf{w}^{(m)}\right). \tag{4.4}$$

The individual loss function $l_{\text{side}}^{(m)}$ for side output m is defined as the balanced cross-entropy loss:

$$\begin{aligned} l_{\text{side}}^{(m)}\left(\mathbf{W}, \mathbf{w}^{(m)}\right) = &- \beta \sum_{j \in Y_+} \log \Pr\left(y_j = 1 | X; \mathbf{W}, \mathbf{w}^{(m)}\right) \\ &- (1 - \beta) \sum_{j \in Y_-} \log \Pr\left(y_j = 0 | X; \mathbf{W}, \mathbf{w}^{(m)}\right), \end{aligned} \tag{4.5}$$

where $\beta = |Y_+| / |Y|$, with $|Y_+|$ denoting the edge ground truth label set. $\Pr\left(y_j = 1 | X; \mathbf{W}, \mathbf{w}^{(m)}\right)$ is computed as the sigmoid function $\sigma(a_j^{(m)}) \in [0, 1]$ on the activation value at pixel j. For each side output layer, this gives an edge map prediction $\hat{Y}_{\text{side}}^{(m)} = \sigma(\hat{A}_{\text{side}}^{(m)})$, where $\hat{A}_{\text{side}}^{(m)} \equiv \{a_j^{(m)}, j = 1, ..., |Y|\}$ are the activations of the side output of layer m.

The side output predictions are combined by adding a weighted fusion layer to the network and simultaneously learning the fusion weight during training. The loss function for this fusion layer is given by:

$$\mathcal{L}_{\text{fuse}}(\mathbf{W}, \mathbf{w}, \mathbf{h}) = \text{Dist}(Y, \hat{Y}_{\text{fuse}})$$

where $\hat{Y}_{\text{fuse}} \equiv \sigma(\sum_{m=1}^{M}) h_m \hat{A}_{\text{side}}^{(m)}$ where $\mathbf{h} = (h_1, ..., h_M)$ is the fusion weight. $\text{Dist}(\cdot, \cdot)$ is the distance between the prediction and ground truth map, which is set as the

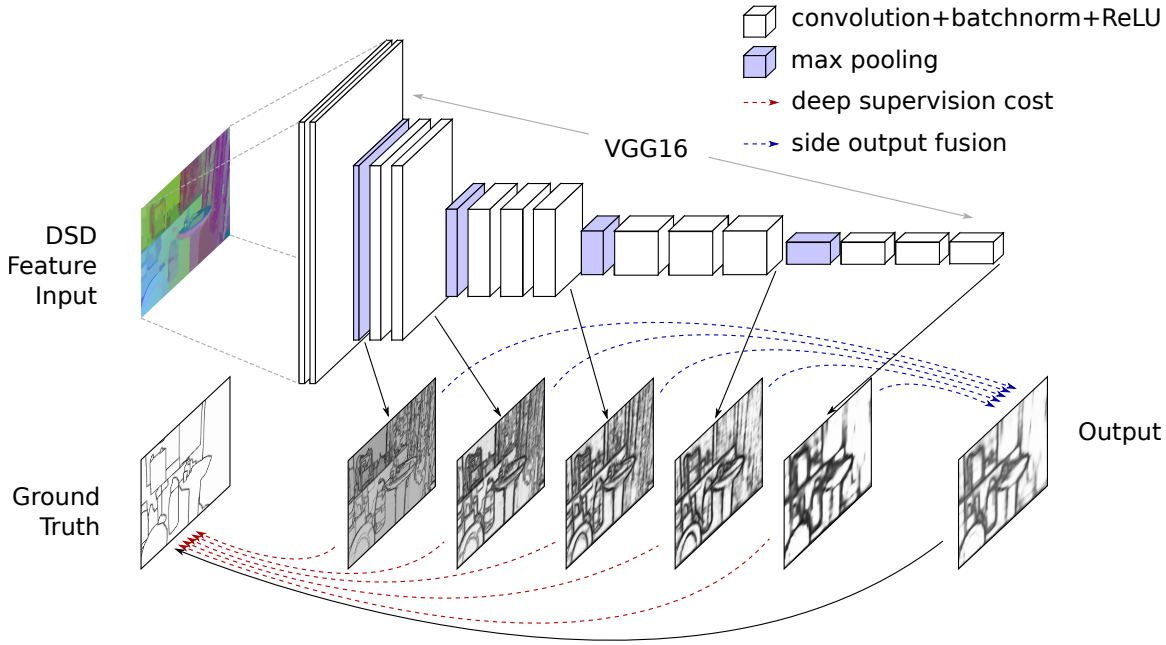


Figure 4-5. An overview of our edge detection system. Our DSD encoding of the depth map is the input to fully convolutional VGG16 network with deep supervision, as in [157]. We add a batch normalization layer after every convolutional layer to speed up convergence.

cross-entropy loss. The combined loss function for the network is thus given by:

$$\mathcal{L}(\mathbf{W}, \mathbf{w}, \mathbf{h}) = \mathcal{L}_{\text{side}}\left(\mathbf{W}, \mathbf{w}\right) + \mathcal{L}_{\text{fuse}}(\mathbf{W}, \mathbf{w}, \mathbf{h})$$

During training this objective function is minimised via standard (back-propagation) stochastic gradient, using batch normalization to speed up convergence. An illustration of the network architecture is shown in Figure 4-5.

We merge the output depth edge maps with rgb maps from the HED architecture [157] in order to assess the contribution of the system as part of an RGB-D edge detector. When merging depth edge with rgb edge maps, we first take the product of the fusion output with all the up-sampled side outputs, since this produces the best results. We observe that the later side outputs produce more semantically meaningful output with some false positives due to blurry edges from up-sampling, whereas the earlier side outputs have excellent edge localization but a high number of false positives due to incorrect edge detections within non-boundary regions. Thus taking the

product of all layers reduces false positives while ensuring that the meaningful edges retain a high response. Multiplying the side outputs in this way increases F-score but decreases average precision. However, when merging with the rgb saliency map, average precision is not reduced.

4.5 Experiments

In this section we detail the implementation of the DSD CNN, and describe the experiments run to evaluate the effectiveness of the encoding.

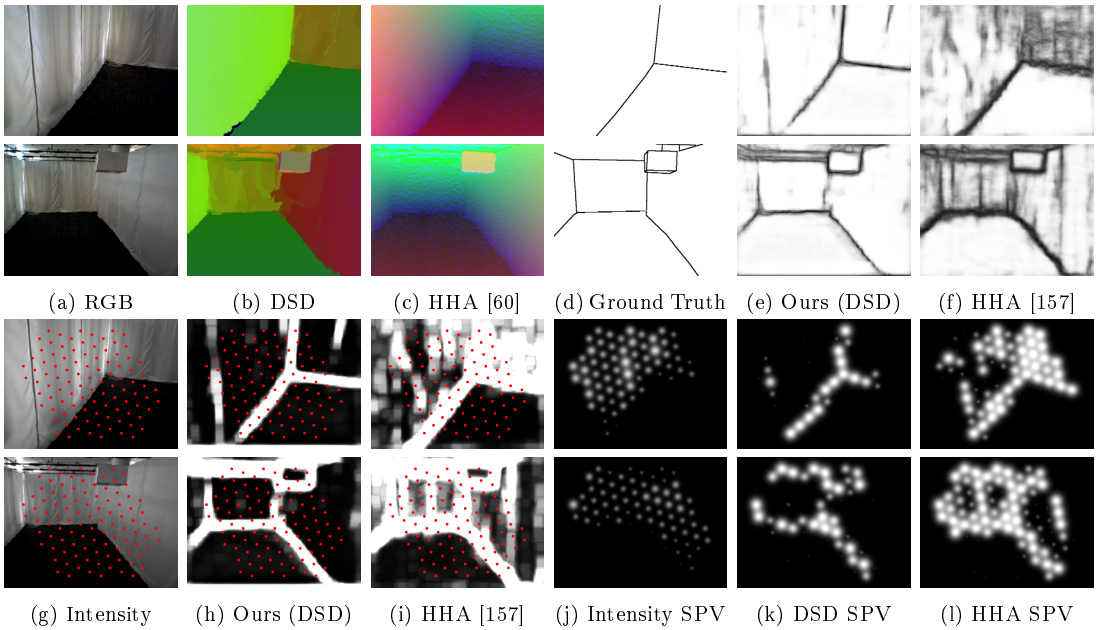
4.5.1 Implementation

We tune the hyper-parameters of the network using the method in [157], using deviations of the F-score on the validation set as a measure of convergence. We select the following hyper-parameter values for our experiments: image size 500×500 mini-batch size = 10, learning rate = $1e5$, momentum = 0.9, weight decay = 0.0002, training iterations = 15000, with learning rate divided by 10 every 5000 iterations.

We fix the coordinate system of the surface normal map $\mathcal{N}$ as follows. We set $\mathbf{u}$ as the inferred direction of gravity and $\mathbf{v}$ as the intersection between the camera plane and the plane defined by $\mathbf{u}$. This provides a stable reference frame for surface orientation measurements, allowing the system to learn extrinsic priors relating to structural edge placement.

4.5.2 Datasets

We evaluate our method on the SUNRGBD dataset, which contains 10335 RGB-D image pairs taken with a variety of commodity depth cameras. As in [43], we convert the segmentation ground truth to edge maps using [59]. Note that SUNRGBD is a superset of the NYU dataset that existing methods use for evaluation, and thus provides a better indication of model performance. We split the SUNRGBD dataset into 6201 training, 2067 validation and 2067 test images.



(a) RGB (b) DSD (c) HHA [60] (d) Ground Truth (e) Ours (DSD) (f) HHA [157]

(g) Intensity (h) Ours (DSD) (i) HHA [157] (j) Intensity SPV (k) DSD SPV (l) HHA SPV

Figure 4-6. (a-f) Challenging examples from our dataset, with output from our method and HHA. Note the ripples in the curtains that produce high local surface normal variation, illustrating the importance of scale for structural edge detection. (g-l) Prosthetic vision inputs and SPV of the scene from intensity, our method, and HED-HHA, with sampling locations shown in red. SPV denotes simulated prosthetic vision.

We also introduce a new dataset, which contains 200 RGB-D image pairs with hand-labeled ground truth. The images were taken with an Asus Xtion Pro depth camera and represent a wide variety of indoor environments, particularly those which would be encountered within robotic grasping or prosthetic vision mobility tasks. Ground truth was provided by a group of volunteers using custom annotation software. Labelers were asked to mark significant structural boundaries in the scene.

We do not perform training on our dataset due to its small size. Rather, we use it to evaluate the generalization ability of the learned edge maps on novel scenes likely to be encountered during mobility tasks.

4.5.3 Comparison with Existing Methods

We compare the surface representation capacity of the DSD feature with the state-of-the-art HHA feature for structural edge detection, by evaluating the quality of learned edge maps on the two features using the HED architecture [157]. Since our

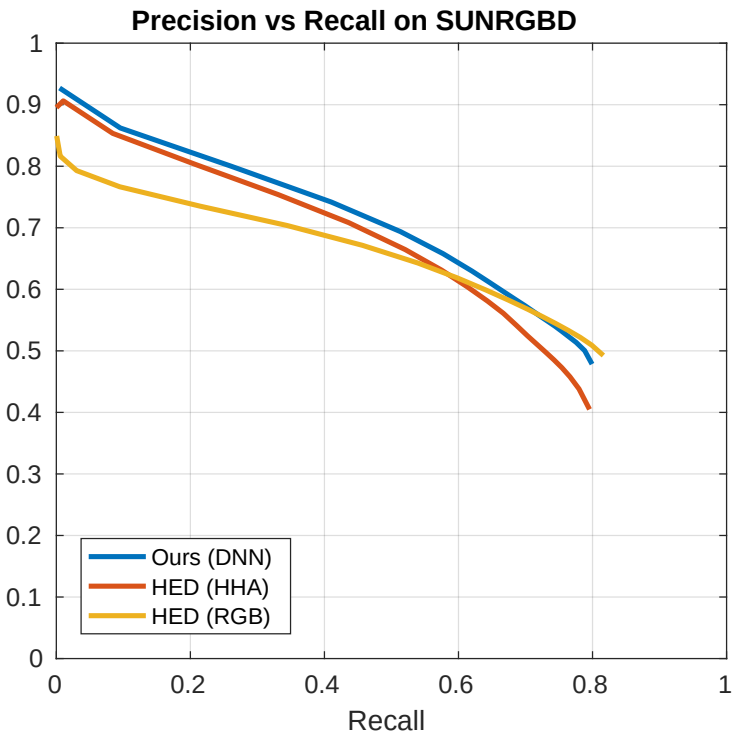
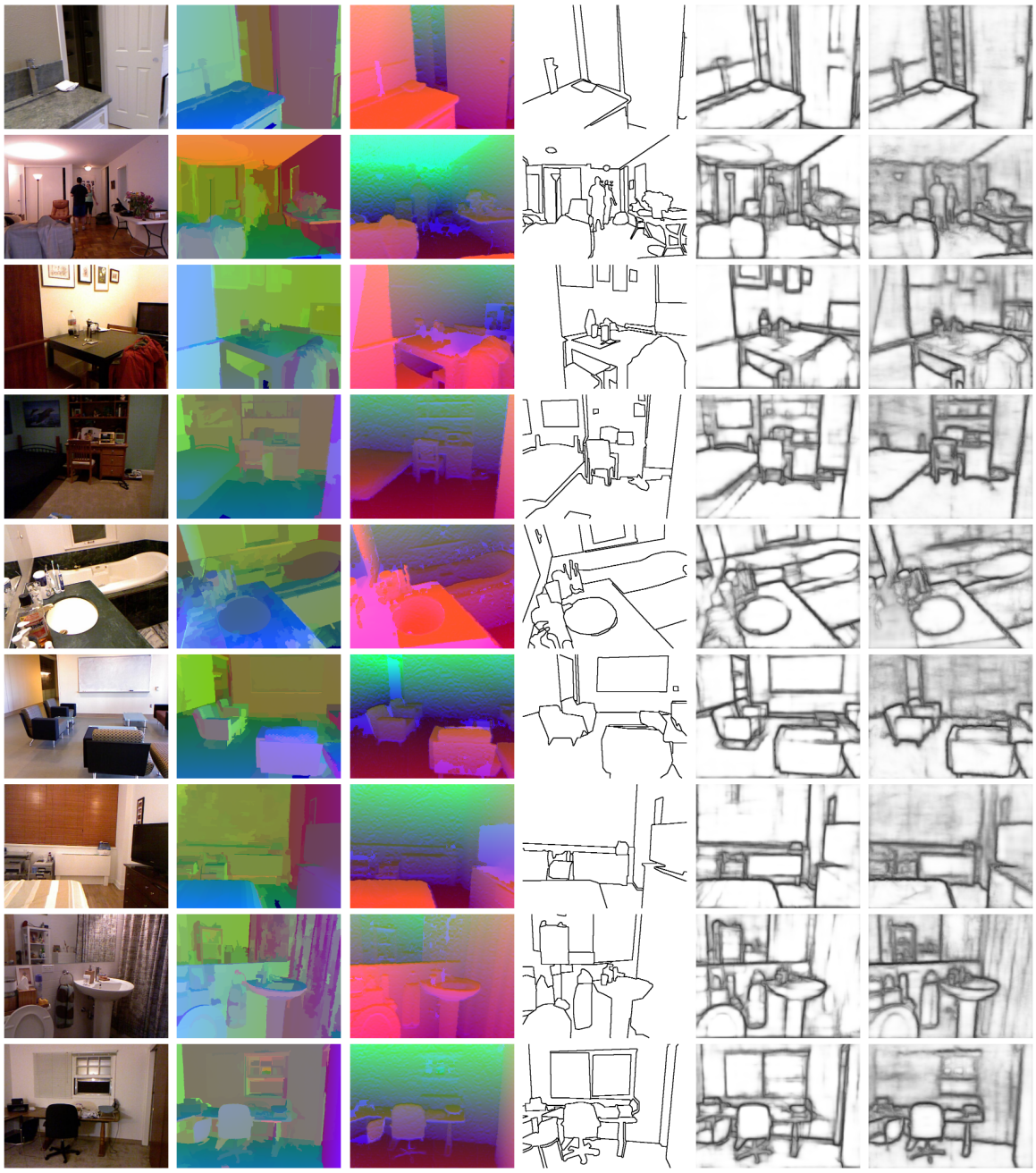


Figure 4-7. PR curve on SUNRGBD.

Method	ODS	OIS	AP
HED HHA	.615	.634	.548
Ours DSD (tuned)	**.630**	**.652**	**.577**
HED RGB	.629	.652	.545
HED RGB+HHA	.649	.672	.606
Ours RGB+DSD	**.652**	**.676**	**.610**

Table 4.3. Results on the SUNRGBD Dataset.

main contribution, the DSD input encoding, is agnostic of the learning framework, we do not provide further comparison with different deep learning architectures. HED is the state-of-the-art base architecture for edge detection, and any optimizations to the framework [94, 81] would likely improve learned results from our feature.



(a) RGB (b) DSD (c) HHA [60] (d) Ground Truth (e) Ours (DSD) (f) HHA[157]

Figure 4-8. Example outputs from the SUNRGBD dataset. Our DSD feature provides a cleaner and more descriptive representation of the boundaries between underlying surfaces, which results in an improved final edge map compared to HHA.

Method	ODS	OIS	AP
HED HHA	.647	.668	.570
Ours DSD	**.678**	**.712**	**.653**
HED RGB	.641	.679	.591
HED RGB+HHA	.679	**.729**	.676
Ours RGB+DSD	**.685**	**.729**	**.685**

Table 4.4. Results on our new depth edge Dataset.

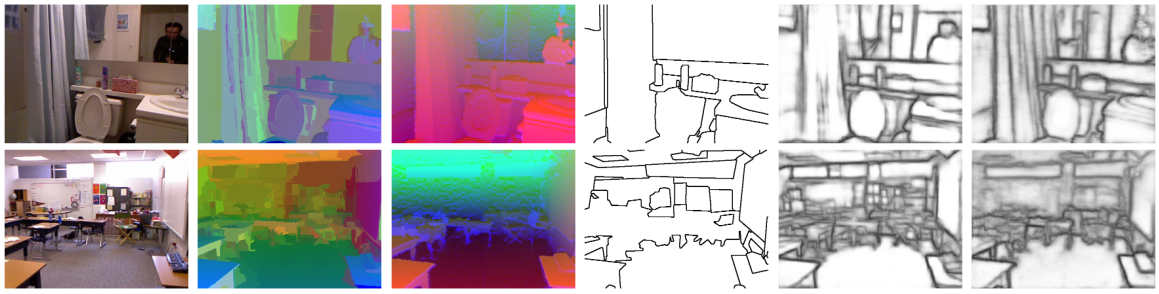
4.6 Results

Our method gives the highest ODS, OIS and AP scores for the depth-only methods, as seen in Table 4.3 and Figure 4-7. This demonstrates that our DSD input encoding makes available to the learning framework a more discriminative surface representation than HHA, enabling more effective classification of edge structure.

From Figure 4-7, we see that while HED-RGB overall doesn't perform as well as depth, it performs relatively well in the high recall region. Thus we see from the graph that the two HED methods can compensate for each other's performance. However, since our curve is mainly above the curve for HED-RGB, there is less potential for performance improvement from a naive combination with HED-RGB. Despite this, our method obtains superior performance when merged with HED-RGB as shown in Table 4.3. Note that the focus of this work is on the depth representation, and further investigation of merging depth and RGB edge maps would increase the RGB-D performance of our system.

To test the cross-dataset generalization of DSD and HHA, we run the pretrained networks on our new RGB-D edge dataset. The results are shown in Table 4.4. Our method outperforms the HHA-based system for depth-only edge detection, demonstrating that the DSD feature provides a more general representation of surface structure.

Figure 4-8 shows some example edge outputs generated by our method. Generally our DSD encoding provides a more effective expression of surface geometry, allowing for a cleaner separation of edge and non-edge structure. For example, in the second last row our method correctly suppresses the depth texture of the shower curtain,



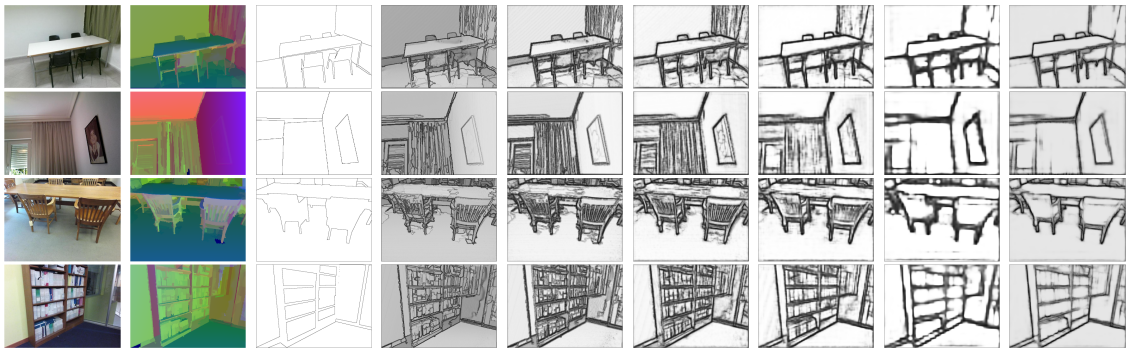
(a) RGB (b) DSD (c) HHA [60] (d) Ground Truth (e) Ours (DSD) (f) HHA[157]

Figure 4-9. Two failure cases for our approach. In the first row our method incorrectly detects salient structure on the mirror. In the second row our method produces a high response on the dense structure in the center of the images, which is not present in the ground truth. Note that these problems also occur in the HHA output.

demonstrating the effectiveness of our encoding when combined with learned high level information.

Figure 4-6 displays simulations of what a prosthetic vision user would perceive when using the standard intensity visual representation, and edge maps produced by our method and HHA on challenging scenes from our new dataset. The errors due to surface normal noise in the HHA SPV can make it difficult for a prosthetic vision user to interpret the scene when performing navigation. Our method reduces noise, providing a clearer depiction of scene structure. The standard intensity scene representation is generally the least informative out of all the SPV images, supporting structure-based depictions of the scene for prosthetic vision.

Failure cases Figure 4-9 shows some examples that represent the typical failure cases of our method. Note that these are also a problem for the HHA system. In particular, the wide range of structural edge labellings spread out over a relatively low number of training examples in the dataset means that it can be difficult to produce an accurate classification in all cases, for example classifying the chairs and tables as non-edge in the second row.



(a) RGB (b) DSD (c) G T (d) Side 1 (e) Side 2 (f) Side 3 (g) Side 4 (h) Side 5 (i) Fusion

Figure 4-10. Visualization of the side outputs produced by the network from DSD input on challenging scenes from SUNRGBD containing structural texture, such as curtains, chair backs, and books. These texture regions are effectively suppressed by the network.

4.6.1 Structural Texture Removal

A major motivation of using a CNN approach was the ability to incorporate local and scene-level context and semantic information to distinguish structural boundaries from structural texture. Figure 4-10 provides a visualization of the deep supervision side outputs produced by our CNN from DSD input on example scenes containing challenging structural texture regions such as curtains, chair backs, and books. The DSD feature provides a high response on these types of regions in the early layers of the network, allowing later layers of the network to effectively suppress these regions. Note that the structural texture has been removed by the system in the final output.

4.7 Chapter Summary

In this chapter we have investigated the extraction of salient edges that support understanding of scene structure for performing physical tasks such as navigation. This is accomplished by learning high level information to help differentiate salient structure, such as room boundaries, from structural texture, such as curtain ripples. We have presented a new depth encoding, the DSD feature, which captures the first order properties of surfaces and facilitates improved classification of surface geometry that corresponds to structural edges.

Our raw DSD feature was compared with the standard HHA depth encoding by applying the Canny edge detector to the respective feature maps. The accuracy of the Canny detector output reflects the suitability of the encoding for edge detection. The DSD feature gave improved performance compared with the HHA feature, demonstrating that our encoding provides a better characterisation of salient edge structure than HHA.

A deep learning model is used to incorporate high level information into our detection system. In our implementation, the DSD feature is incorporated into a fully convolutional neural network adapted from the VGG16 network and trained using deep supervision. Our system achieves state-of-the-art structural edge detection results a large scale existing dataset. Furthermore, since the proposed DSD feature is agnostic to the edge detection method, it can be combined with many existing edge detection frameworks. The more effective encoding of scene structure would improve the accuracy of edge detection in these systems.

Our proposed edge detection system also outperforms existing methods on our new RGB-D edge dataset which contains a range of navigation scenarios for prosthetic vision. Our method is able to reduce false positives from noisy depth measurements and structural texture, making it easier to interpret the salient structure in a prosthetic vision scenario. Through our approach and the results, this chapter thus addresses the thesis subproblem of identifying salient edges that convey scene shape for prosthetic vision.

Chapter 5

Surface Orientation for Salient Object Detection

The preceding two chapters have investigated the thesis subproblem of detecting salient edges for conveying scene shape. In Chapters 5 and 6, we now shift the focus onto the determination of what kind of general object structure is salient to the human visual system, for performing visual attention direction on a prosthetic vision display.

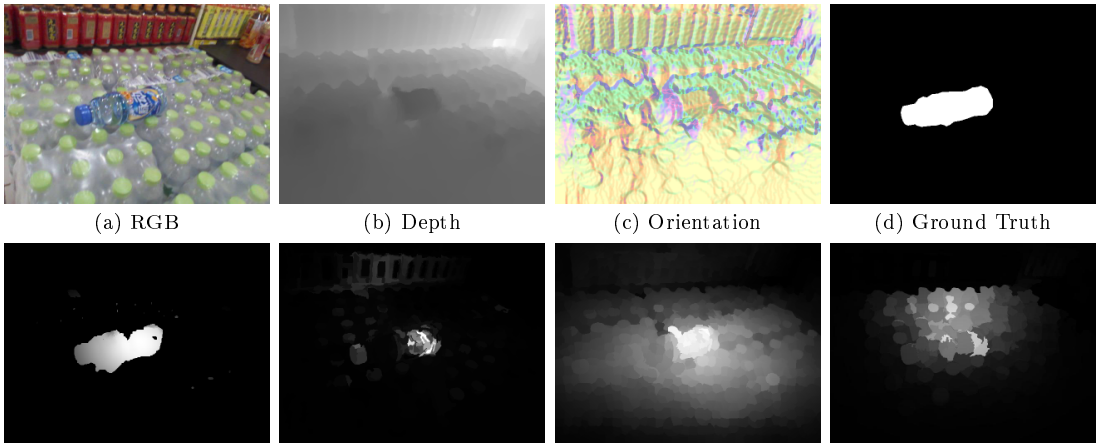
This chapter investigates the extraction of structurally salient objects from the scene, based on analysis of surface orientation distribution contrast. We first introduce the histogram of surface orientations (HOSO) feature in Section 5.2, which uses the first order properties of the depth image for surface representation. This feature is incorporated into a salient object detection system, described in Section 5.3. Experiments and results on two RGB-D salient object detection datasets are presented in Sections 5.4 and 5.5 respectively. The chapter is concluded in Section 5.6.

5.1 Introduction

In this chapter we aim to identify what types of object structure are salient to the human visual system. Knowledge of structurally salient objects and regions in a scene is crucial for performing many tasks of everyday living with prosthetic vision, such as tabletop tasks, grasping tasks, orientation within the environment, and obstacle

D	input depth map
I	input RGB map
$\mathrm{dom}(D)$	set of pixel positions
$(\theta_u(x,y), \theta_v(x,y))$	2D orientation of (x,y)
$\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$	principal components obtained using PCA
b	2D histogram binning function
P	image patch
$H(P)$	patch histogram
$D(P)$	mean patch depth
$I_{\mathrm{lab}}(P)$	mean patch LAB colour
$\mathcal{P}$	set of patches from image segmentation
$N_{P,r}$	neighbourhood set of patches within radius r of P
$\mathcal{N}_{P,R}$	scale space of P defined by set of radii R
dist_B	Bhattacharyya distance
card	cardinality of a set
$k_h, k_d, k_{\mathrm{lab}}$	KDE kernel components for HOSO, depth, and LAB
$\sigma_h, \sigma_d, \sigma_{\mathrm{lab}}$	KDE bandwidths for HOSO, depth, and LAB
C_{hoso}	contrast function
S_{hoso}	low-level saliency
S_{prior}	prior adjusted saliency
(μ_x, μ_y, μ_d)	estimated object center
$(\sigma_x, \sigma_y, \sigma_z)$	estimated object size
S_{obj}	estimated salient object map
M_{gc}	graph cut refinement mask
$\mathcal{S}_{\mathrm{hoso}}$	final saliency output

Table 5.1. List of symbols used in this chapter.



(a) RGB (b) Depth (c) Orientation (d) Ground Truth

(e) Our Method (f) GP [121] (g) ACSD [73] (h) LMH [113]

Figure 5-1. Saliency output on an image with low foreground depth contrast. Our method measures surface orientation distribution contrast to effectively identify foreground structure. Output is shown for three state-of-the-art methods Global Priors [121], Anisotropic Center Surround Difference [73], and Low Medium High [113].

avoidance. Detection of these regions enables vision processing methods to direct attention towards important parts of the scene in a manner that resembles the deployment of biological visual attention. We perform detection of these salient regions by analysing surface shape as captured by surface orientation distribution contrast.

We make the observation that, in terms of depth, saliency consists of not just how close an object is, but that it has an unusual profile of surface orientation with respect to its local region or to other parts of the scene, or has an overall orientation that is unusual. For example the corner between a wall and floor, an obstacle on the ground, or clutter in a tidy space. Surface orientation contrast thus offers a promising structural measure of saliency at multiple scales that operates independently to depth, and can be used to complement depth-based contrast. However, while first order surface properties are commonly used for tasks such as 3D object recognition [143], incorporation of surface orientation for saliency detection has received much less attention [35, 42, 121].

In this chapter, we present a new unified model for salient object detection that integrates surface orientation, depth, and color contrast at multiple scales. Unlike previous approaches, we integrate both orientation and depth contrast in a consistent framework, taking advantage of the complementary information they offer. Surface

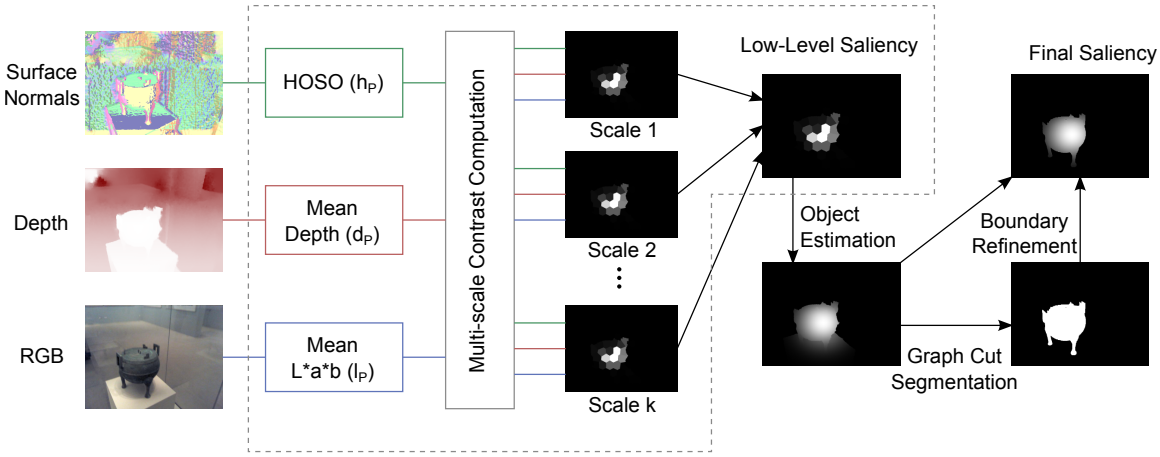


Figure 5-2. Overview of the main stages of our method. We measure multi-scale contrast of orientation, depth, and colour to obtain a low-level saliency map. We use the low-level saliency to estimate and object map, and then perform boundary refinement using a graph cut approach.

orientation contrast in existing methods is computed only at a global scale [42] or only with respect to similar regions in the image [35], which can lead to an increased number of false positives and false negatives respectively. Instead, our unified model performs a multi-scale measurement of orientation contrast, based on the intuition that salient objects are likely to remain distinct across multiple scales. Unlike purely global formulations of surface orientation, our method captures locally unusual surface orientation profiles that characterize many types of structurally interesting regions, such as wall-floor edge boundaries. Furthermore, while previous work represents regions using mean orientation, we introduce the histogram of surface orientation (HOSO) feature for RGB-D saliency to capture the distribution of surface normals, providing a robust and descriptive characterisation of the underlying region. While histogram based representations of first order image properties are common in feature detection and matching [40], their use and effectiveness is unexplored for RGB-D salient object detection.

Contrast computation in our system is performed using a Gaussian KDE [113]. This allows the integration of different feature types during computation, rather than fusing individually computed feature contrast maps [110, 48], to better exploit the strong complementarities between surface orientation, depth, and color. The incor-

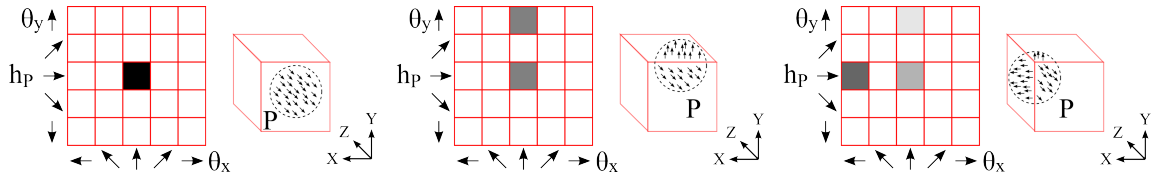


Figure 5-3. Illustration of the HOSO feature for three different patches on a cube, with camera direction along the Z axis. Given an image patch P, surface normals within P are parameterized by their 2D orientation and binned into a 5×5 histogram h_P.

poration of multiple discriminative features tends to produce a precise but sparse low-level saliency map. We post-process this map using object map estimation and boundary refinement procedures to obtain a uniform saliency response across detected objects. We evaluate our model on two recently proposed RGB-D datasets for salient object detection, achieving superior performance to existing state-of-the-art methods at the time of contribution. Furthermore, we demonstrate the contribution of each feature type and computation stage to the overall performance of our model.

The main contributions of this work are: insight that surface orientation distribution contrast provides valuable cues for determining locally unusual structure that is indicative of salient objects, and a novel feature, HOSO, for capturing these cues; proposal of the first unified multi-scale saliency detection system incorporating surface orientation, colour, and depth contrast; and demonstration of the effectiveness of HOSO and our system through state-of-the-art results at the time of contribution on two datasets.

5.2 HOSO Feature

Our saliency model includes the distribution of surface orientation as a feature, based on the observation that salient objects are more likely to contain surface orientation structure that contrasts with the surroundings.

We aim to identify structurally salient regions based on their surface orientation profile. In order to perform this task, the representation of patch-level surface orientation must be descriptive as well as robust to noise. First-order surface properties are

particularly sensitive to sensor noise, which can impact the performance of a saliency system if used directly [110].

Rather than representing a patch with a single orientation value as in previous work, we use a histogram to capture the distribution of patch normals as the core orientation feature. This provides a more detailed representation of the underlying surface shape, and improves the capacity of the feature for distinguishing locally unusual structure. Furthermore, histograms are more robust to sensor noise than mean values.

The HOSO feature is computed as follows. First, point-wise normals are estimated from the input depth image D using PCA. Given an image point (x, y) in the input, the PCA support region is defined as the set of nearby points within a distance of 5 pixels. This support size was found to be large enough to alleviate the effect of noise, while not being so large as to produce overly-smoothed normals. Each point from this support region is projected to a 3D vector representing the spatial position of the point in the camera coordinate system, and PCA is performed on the resulting set of 3D points in order to obtain the three principal components $\mathbf{V_1}, \mathbf{V_2}, \mathbf{V_3} \in \mathbb{R}^3$, ordered according to decreasing eigenvalue. $\mathbf{V_1}$ and $\mathbf{V_2}$ thus span the plane that fits the support region with least squared error, and $\mathbf{V_3}$ is the desired normal vector for the plane. We parameterise the fitted normal $\mathbf{V_3}$ by its 2D orientation to avoid wrap around issues and facilitate uniform quantization. The 2D normal orientation of (x, y) is defined as in Equation 4.1 and denoted as $(\theta_u(x, y),\ \theta_v(x, y))$.

Normal orientations in an image patch P are binned into a normalized 2D histogram $H(P)$ (see Figure 5-3). Setting the size of $H(P)$ to be 5 bins in each dimension was found to achieve a good balance between descriptiveness, robustness, and efficiency for HOSO. The bin mapping function of $H(P)$ is given by

$$b(x, y) = \left(\left\lfloor 5 \cdot \frac{\theta_u(x, y)}{\pi} \right\rfloor, \left\lfloor 5 \cdot \frac{\theta_v(x, y)}{\pi} \right\rfloor \right). \tag{5.1}$$

Thus, each point $(x, y) \in P$ increments bin $b(x, y)$ of $H(P)$. The value of bin (i, j) of

$H(P)$ is thus given by

$$H_{i,j}(P) = \frac{\text{card}\left(\{(x,y) \in P \mid b\,(x,y) = (i,j)\}\right)}{\text{card}(P)}. \tag{5.2}$$

The dissimilarity of the surface orientation distributions of two patches is measured using the Bhattacharyya distance $\text{dist}_B(\cdot,\cdot)$ between their HOSO features:

$$\text{dist}\,(P,Q) = \text{dist}_B\,(H(P), H(Q))\,. \tag{5.3}$$

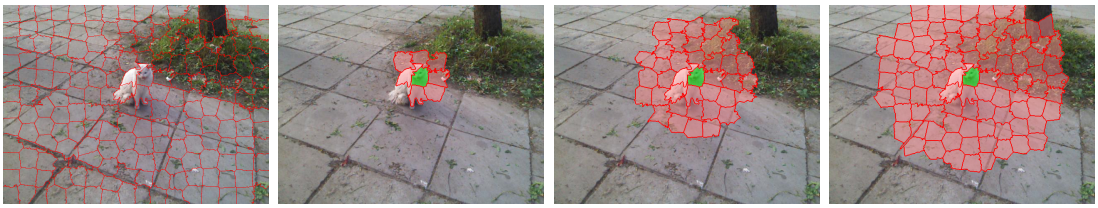
The Bhattacharyya distance is a standard metric for measuring the difference between discrete distributions in statistics, and is commonly used to perform histogram comparison in computer vision applications [37, 55, 122, 34]. This metric takes into account not only the means but also the variances of the distributions being compared, allowing for improved discrimination of different surface structure.

5.2.1 Patch-level Feature

Given an image patch P, we aim to compute saliency based on contrast from 3D structure as well as appearance. We use the HOSO feature $H(P)$ to capture surface shape, and in addition mean depth $D(P)$ to capture relative position in the scene and mean LAB colour $I_{\text{lab}}(P)$, obtained from converting the mean RGB colour $I(P)$, to represent appearance [113]. Thus $H(P)$, $D(P)$, and $I_{\text{lab}}(P)$ form our patch representation and the basis for contrast computation.

5.3 Saliency Detection System

The pipeline of our method consists of three major stages, as shown in Figure 6-5. First, a low-level saliency map is computed from surface orientation, depth, and colour contrast at multiple scales. We use Gaussian Kernel Density Estimation [113] to measure contrast and integrate the different features during the contrast computation stage. This is followed by an object estimation stage, which uniformly highlights



(a) Segmentation (b) $r = 60$ (c) $r = 140$ (d) $r = 220$

Figure 5-4. Example image segmentation and illustration of contexts at multiple scales. The candidate patch P is shown in green. The context sets $N_{P,r}$ are shown in red, containing patches within distance r of P.

foreground regions identified in the low-level saliency map. Each pixel is assigned a probability that it belongs to the foreground, computed using a Gaussian model of the object constructed from the low-level saliency map. In the final step, the boundaries of the estimated object map are refined with a graph cut based approach [105].

5.3.1 Low-level Saliency

This section describes our method for computing the low-level saliency map from raw patch level features. We first segment the input image into a set of patches $\mathcal{P}$ using SLIC [2]. The low-level saliency score of a patch $P \in \mathcal{P}$ is based on its contrast with a set of neighbouring patches N. The contrast is measured by estimating the probability $\text{prob}(P|N)$ that P comes from the distribution defined by N in feature space, as in [113]. A low value of $\text{prob}(P|N)$ implies that P is unlikely to belong to N, and has a high contrast.

We use a kernel density estimator to compute $\text{prob}(P|N)$ [113]. However, in addition to mean depth and colour, we extend the density estimation to include the HOSO feature, incorporating differences of surface orientation distributions into the density function. If a patch has an unusual surface orientation profile compared to its surroundings, such as a ball resting on the ground, then it will have a low estimated probability of being part of the context distribution, and consequently a high saliency score. On the other hand if a patch has an almost identical surface orientation profile to its surroundings, such as a patch on a planar surface, then the estimated probability density function will have a high value at the HOSO feature of the point, leading to

a low saliency score.

The probability density estimation is thus given by:

$$\text{prob}\left(P|N\right) = \frac{1}{\text{card}(N)} \sum_{Q \in N} k_h\left(H(P), H(Q)\right) k_d\left(D(P), D(Q)\right) k_{\text{lab}}\left(I_{\text{lab}}(P), I_{\text{lab}}(Q)\right), \tag{5.4}$$

where $k_h(\cdot,\cdot)$, $k_d(\cdot,\cdot)$, and $k_{\text{lab}}(\cdot,\cdot)$ are the kernel components corresponding to surface orientation distribution, mean depth, and mean colour respectively.

We define the surface orientation distribution component as a Gaussian kernel with bandwidth $\sigma_{h,P,Q}$:

$$k_h(H(P), H(Q)) = \exp\left(-\frac{\text{dist}_B\left(H(P), H(Q)\right)^2}{2\sigma_{h,P,Q}^2}\right). \tag{5.5}$$

The estimate is obtained by measuring the Bhattacharyya distance from the HOSO feature $H(P)$ of the candidate patch to the density function of the neighbour patch Q.

As in [32] we observe that objects that are closer than their surroundings are more likely to be salient. We aim to limit the contribution of context patches with lower depth than the candidate patch, since in these cases the candidate patch is more likely to be background. This is achieved by scaling the base KDE bandwidth σ_h, depending on whether the candidate patch is in front of the context patch:

$$\sigma_{h,P,Q} = \begin{cases} \sigma_h & \text{if } D(P) > D(Q) \\ 2\sigma_h & \text{otherwise.} \end{cases} \tag{5.6}$$

This increases the bandwidth and reduces the influence of the neighbour patch if it is further away than P.

The depth and colour Gaussian kernels $k_d(\cdot,\cdot)$ and $k_{\text{lab}}(\cdot,\cdot)$ are defined similarly, using the Euclidean distance between feature values instead of Bhattacharya distance, and with respective bandwidths $\sigma_{d,P,Q}$ and $\sigma_{\text{lab},P,Q}$.

The contrast measurement function is thus given by:

$$C_{\text{hoso}}(P, N) = -\log\left(\text{prob}\left(P|N\right)\right). \tag{5.7}$$

The low-level saliency $S_{\text{hoso}}(P)$ of a patch P is formulated as the product of the contrast measurement function across multiple scales, such that:

$$S_{\text{hoso}}(P) = \prod_{N \in \mathcal{N}_P} C_{\text{hoso}}(P, N), \tag{5.8}$$

where $\mathcal{N}_{P,R} = \{N_{P,r} | r \in R \subset \mathbb{R}\}$ denotes the scale space of P, and the neighbourhood $N_{P,r}$ of P consists of all other patches within a radius r of P (see Figure 5-4). That is, $N_{P,r} = \{Q | \|(x_P, y_P) - (x_Q, y_Q)\|_2 < r \text{ and } Q \neq P\}$, where (x_P, y_P) and (x_Q, y_Q) are patch centroids.

The contribution of each feature when computing low-level saliency on RGBD-1000 is shown in Figure 5-5. Note that incorporating orientation with depth results in a larger improvement than using colour and depth, validating the incorporation of surface orientation as a structural saliency feature. The combination of all three features gives the best performance, indicating that each feature contributes positively to the final result.

5.3.2 Priors

We apply two standard saliency priors to reweight the low-level saliency map S_{hoso}. The first is the depth prior, which incorporates information about absolute depth into the system. The depth prior is applied by dividing patch saliency with mean depth. The second prior is the spatial prior, which reweights patch saliency according to a distance from the center of the based on a Gaussian distribution [99]. Application of these saliency priors is a standard step in RGB-D saliency systems. See Chapter 2 for further information about these saliency priors. We denote the prior adjusted saliency map as S_{prior}.

5.3.3 Salient Object Map Estimation

The low-level saliency computation stage tends to produce saliency maps characterized by sparse high-saliency patches. The multiplicative aggregation of complementary and discriminative features can result in a low overall saliency score for a patch if one feature is assigned a low contrast. Thus, only a few highly distinct points produce a high saliency score in the low-level map.

Ensuring a consistently strong saliency response across entire objects is a fundamental objective in salient object detection [19]. We use the low-level saliency map described in Section 5.3.1 to build a Gaussian model of the object based on image position and depth, from which each pixel is assigned a score reflecting the probability that it is part of the salient object. This is implemented in a similar way to the high-level object bias enhancement performed in [113], but with mean and variance computation modified to account for a saliency map with sparse regions of high response.

The probability that a pixel (x, y) in the input depth image belongs to a salient object is computed based on the estimated location and size of the object. The estimated object map S_{obj} is obtained using a Gaussian model, given by:

$$S_{\text{obj}}(x, y) = \exp\left[-\left(\frac{x-\mu_x}{2\sigma_x}\right)^2 - \left(\frac{y-\mu_y}{2\sigma_y}\right)^2 - \left(\frac{D(x,y)-\mu_d}{2\sigma_d}\right)^2\right], \tag{5.9}$$

where (μ_x, μ_y, μ_d) is the expected object center, and $(\sigma_x, \sigma_y, \sigma_d)$ is the expected object size.

We will now detail the computation of (μ_x, μ_y, μ_d) and $(\sigma_x, \sigma_y, \sigma_d)$. Let $S_{\text{prior}}(x, y)$ denote the prior adjusted saliency of a pixel (x, y), obtained by propagating patch saliency to member pixels. In order to handle a saliency map with sparse regions of high response, we set the expected object center as the weighted mean over all pixels:

$$(\mu_x, \mu_y, \mu_d) = \frac{\sum_{(x,y)\in \text{dom}(D)} S_{\text{prior}}(x, y)(x, y, D(x, y))}{\sum_{(x,y)\in \text{dom}(D)} S_{\text{prior}}(x, y)}. \tag{5.10}$$

The expected object size is based on the weighted variance of the image:

$$\left({\sigma_x}^2, {\sigma_y}^2, {\sigma_d}^2\right) = \frac{\sum_{(x,y)\in \mathrm{dom}(D)} S_{\mathrm{prior}}(x,y)\left((x-\mu_x)^2, (y-\mu_y)^2, (D(x,y)-\mu_d)^2\right)}{\sum_{(x,y)\in \mathrm{dom}(D)} S_{\mathrm{prior}}(x,y)}. \tag{5.11}$$

Since low-level saliency may not be high at all the extremities of the object, we set the estimate of expected object size to three standard deviations, by scaling $(\sigma_x, \sigma_y, \sigma_d)$ by three.

5.3.4 Boundary Refinement

The estimated object map H from the previous stage may contain inaccurate foreground boundaries, particularly when the object occupies a similar depth range to nearby background. Boundary refinement is a common post-processing step employed in existing salient object detection systems (e.g. [32, 113, 105]). We use the graph cut based saliency refinement method described by [105] to obtain object boundaries based on appearance information. The refinement process iteratively builds a model of the foreground and background based on initial seed regions. We provide the thresholded object saliency map S_{obj} as the binary foreground seed image for the graph cut refinement process. A threshold of 0.8 is used to ensure only high confidence saliency regions form the basis of foreground model estimate. The graph cut segmentation process produces an output binary mask M_{gc} that denotes the set of pixels that form the predicted object, and is used to prune non-foreground areas from S_{obj}. The final pixel-wise saliency is thus given by

$$\mathcal{S}_{\mathrm{hoso}}(x,y) = M_{\mathrm{gc}}(x,y) \cdot S_{\mathrm{obj}}(x,y). \tag{5.12}$$

5.4 Experiments

We evaluate our method on two recently proposed datasets for RGB-D salient object detection. The first is RGBD1000 [113], which was introduced to address the lack of a large dataset with depth information for salient object detection. It contains

1000 images featuring diverse scene and object types, with low depth and colour contrast between the foreground and background. We also report the performance of our method on the NJUDS2000 salient object detection dataset [73], containing 2000 disparity images computed from stereo image pairs.

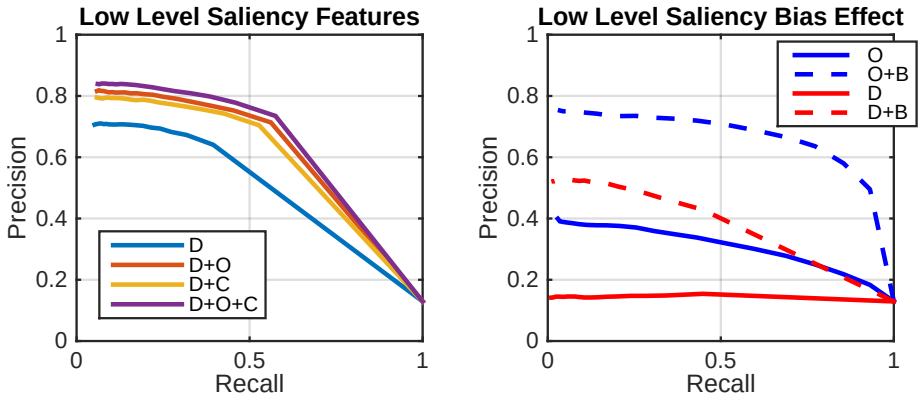
Our method is compared with three state-of-the-art RGB-D salient object detection systems: Low-Medium-High Saliency (LMH) [113] proposed by the authors the RGBD1000 dataset, Anisotropic Center Surround Difference (ACSD) [73], from the authors of the NJUDS2000 dataset, and Global Prior saliency (GP) [121]. We also include comparisons to two top ranking 2D saliency algorithms according to a recent survey [19]: DSR [88], and DRFI [70].

We examine the effect of center and depth bias on low-level orientation contrast saliency compared to the low-level depth saliency from [113]. Note that since [113] is only available as a single executable, we use a custom implementation of the low-level saliency which omits center and depth prior application. We also measure the performance of the low-level and object estimation stages of our framework, and examine the contribution of the different feature types used in our low-level saliency computation method.

5.4.1 Contrast Computation Scales

We perform an analysis of structural feature contrast at different scales for foreground identification on the dataset, in order to help inform scale selection for our saliency system.

Figure 5-6 shows that for a small scale size, foreground patches typically have higher contrast with orientation than depth. In particular, a large number of foreground patches have low local depth contrast, suggesting that depth contrast provides poor discriminability at a local scale, and that orientation contrast is more likely to distinguish foreground regions when the context size is small. However, background regions tend to have greater orientation contrast for larger scales than depth contrast, suggesting that the former is not suited for large context sizes. Based on these observations, we omit depth and orientation when computing contrast with small and



(a) (b)

Figure 5-5. (a) Comparison of low-level saliency results on RGBD1000 when incorporating various patch feature combinations. D=mean depth, O=surface orientation histogram (HOSO), C=mean LAB colour. This shows that the effect of surface orientation is large if there is a lack of colour information, for example in a low contrast environment or under low lighting conditions. In the case that colour is incorporated, using HOSO provides additional improvement. (b) The effect of center and depth bias on low-level saliency using our orientation feature (O) and a custom implementation of the low-level depth saliency term of DCS (D).

large context sizes respectively.

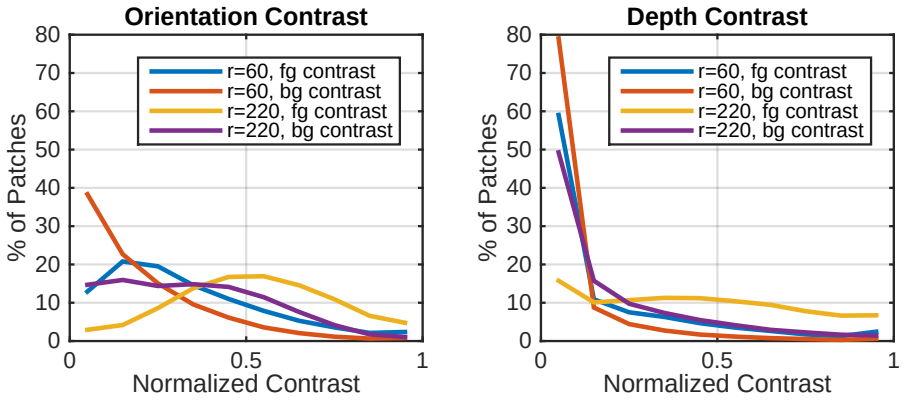
5.4.2 Implementation Details

In the experiments, we measure contrast across three scales, $R = \{60, 140, 220\}$. These scales were selected to produce small, medium, and large neighbourhoods for each patch, produce good results for general use. The KDE bandwidths in Equation 5.4 of the mean depth and Lab colour features were set to $\sigma_d^2 = \sigma_l^2 = 0.025$. For orientation, bandwidths of $\sigma_h^2 = 0.1$ for scale 60 and $\sigma_h^2 = 0.3$ for scale 140 were found to work well.

Our unoptimized implementation takes approximately 7 seconds per 640×480 image running on a 2.6GHz i5 processor with 8GB of RAM.

5.4.3 Evaluation Metrics

Performance is evaluated through the precision-recall curve and mean F-score, the F_β measure with $\beta = 0.3$ emphasizing precision [4]. The F-score is computed from the



(a) (b)

Figure 5-6. Analysis of contrast for (a) surface orientation and (b) mean depth features at multiple scales on RGBD1000, displaying the percentage of foreground (fg) and background (bg) patches that exhibit the normalized contrast values with respect to a neighbourhood of radius r.

saliency output using an adaptive threshold equal to twice the mean of the image [4].

5.5 Results

Our method produces a superior F-score compared to all other methods on both datasets, as seen in Figures 5-7c and 5-7d. Furthermore, our method achieves a consistently high performance across the two datasets whereas the other methods tend to favour one or the other.

Figure 5-7a shows that our system achieves higher precision than other methods at comparable recall rates on RGBD1000. The increased precision is most apparent at just under 0.8 recall. At this point our method is able to identify a larger portion of foreground regions than other methods without affecting precision. Similarly, Figure 5-7b shows that our method has the highest precision up to just under 0.7 recall. Figures 5-7a and 5-7b also show the contribution of each computation stage in our framework. We see from the figure that applying the object estimation map significantly improves results compared to the low-level saliency map, in particular boosting recall as we expect. The application of boundary refinement subsequently increases the precision of the estimated object map. This pattern of improvement

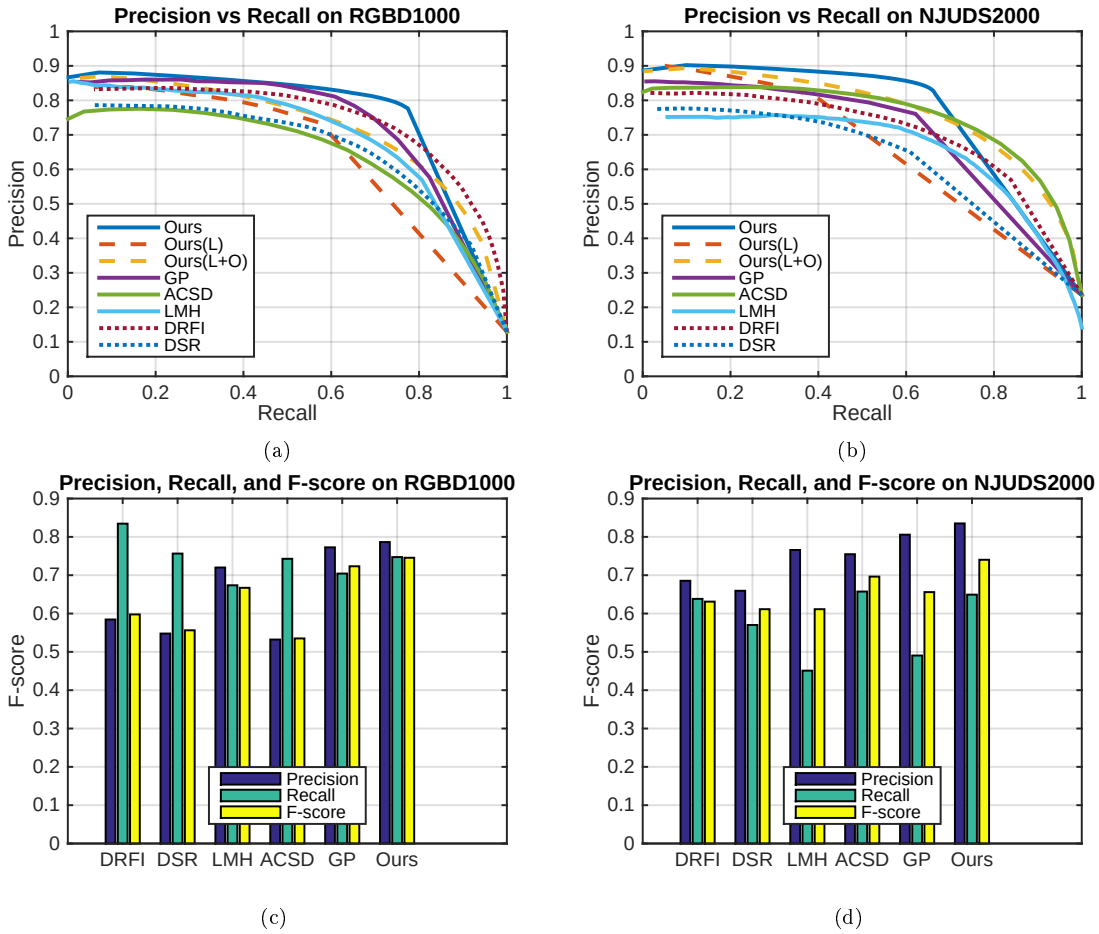


Figure 5-7. Quantitative comparisons of performance over RGBD1000 and NJUDS2000 datasets. Ours(L) denotes our low level saliency map, and Ours(L+O) denotes our estimated object map.

follows the aim of each stage: identification of salient regions, expansion of candidate regions to cover foreground objects, and boundary refinement to remove background.

We plot the precision-recall for our low-level saliency map using different feature combinations in Figure 5-5a. As expected, using individual features gives relatively low scores. Combining depth and orientation produces better results than combining depth and colour, which highlights the complementary nature of the two structural features. The relatively high performance of this pairing suggests that orientation may be used as an alternative when colour is not available or reliable, such as in environments with low contrast or poor lighting. The best performance is observed when using all three feature types, demonstrating that each feature offers distinct

information that is extracted effectively in our framework. As shown in Figure 5-5b, the low-level surface orientation saliency of our method outperforms the low-level depth saliency of [113] both with and without the bias terms. This demonstrates that surface orientation contrast is a more reliable indicator of foreground than depth contrast, particularly near image boundaries.

5.6 Chapter Summary

In this chapter, we have addressed the thesis subproblem of identifying salient object structure for emulating biological visual attention deployment. We have found that regions with irregular surface shape compared to their surroundings are likely to occur on salient objects, and that surface orientation distribution contrast provides useful information for determining these regions. We thus introduce the HOSO feature to measure surface orientation distribution contrast for RGB-D saliency.

We propose a new unified model that integrates surface orientation distribution contrast with depth and color contrast across multiple scales. This model is implemented in a multi-stage saliency computation approach that first performs low-level contrast estimation using a kernel density estimator (KDE), and then performs the post-processing steps of applying priors, estimating object positions from the low-level saliency map, and refining the estimated object positions using graph cut. The HOSO feature increases the precision of low-level saliency predictions, allowing for more effective salient structure segmentation during post-processing.

Our method is evaluated on two RGB-D salient object detection databases, achieving superior performance to previous state-of-the-art methods. Furthermore, our method achieves a consistently high performance across the two datasets whereas the other methods tend to favour one or the other. This demonstrates that surface shape, as represented by the HOSO feature, plays an important role in determining structural saliency.

Based on these findings, we conclude that use of the HOSO feature results in improved capacity to detect objects that are salient to the human visual system.

Therefore, use of surface orientation distribution contrast would benefit vision processing methods that aim to perform attention direction towards salient objects in the environment.

(a) RGB (b) Depth (c) Ori (d) G. T. (e) Ours (f) [121] (g) [73] (h) [113]

Figure 5-8. Comparison of saliency maps generated by state-of-the-art systems. Our method is shown with GP [121], ACSD [73], and LMH [113]. G. T. denotes Ground Truth and Ori shows surface orientation.

Chapter 6

Structural Saliency from Local Background Enclosure

This chapter continues the investigation into the thesis subproblem of identifying the types of structure that are salient to the human visual system, and presents a new depth feature called local background enclosure (LBE) for structural salient object detection. The LBE feature addresses fundamental limitations of contrast-based depth methods such as the system presented in Chapter 5; that saliency is often skewed by background regions of high contrast, and that depth contrast can vary significantly depending on viewpoint and the exact distances between objects. This work offers further insight into what kind of object structure is salient to humans, which can be used to better inform prosthetic vision scene representations on object-related tasks such as obstacle avoidance and tabletop tasks. The proposed LBE feature is described in Section 6.2, and Section 6.3 describes our salient object detection system. Sections 6.4 and 6.5 describe the experiments and results evaluating both the raw LBE feature as well as the saliency detection system on two RGB-D salient object detection datasets. Section 6.6 summarises the work presented in the chapter, and Section6.7 concludes the technical portion of the thesis.

D	input depth map
I	input RGB map
$\mathcal{P}$	set of patches obtained from image segmentation
N_P	neighbourhood of P
$\mathcal{B}(P,t)$	background set of P with depth difference threshold t
η	background intersection indicator function
Θ	set of boundaries of angular regions not containing background
f, g	angular fill and gap density functions
F, G	angular fill and gap distribution functions
S_{lbe}	low-level lbe saliency
S_{prior}	prior adjusted saliency
$\mathcal{S}_{\text{lbe}}$	final saliency output

Table 6.1. List of symbols used in this chapter.

6.1 Introduction

This chapter aims to improve salient object detection from depth input. While the method in Chapter 5 accomplishes this to some degree, methods that operate on depth contrast have two major weaknesses for application in prosthetic vision.

- Depth contrast methods are prone to false positives and false negatives due to background regions with large depth difference. Figure 6-2 shows an example in which the foreground has relatively low depth contrast, making it challenging to detect using existing depth features. Contrast in background regions is generally unavoidable, and can lead to false detections that impinge the ability of a prosthetic vision scene representation to facilitate safe navigation.

- Depth contrast magnitude does not reflect a consistent and meaningful quantity with respect to the needs of prosthetic vision, and in general depth contrast in a scene is largely dependent on random factors such as object placement and viewpoint. For example, Figure 6-1 shows a scenario where two identical

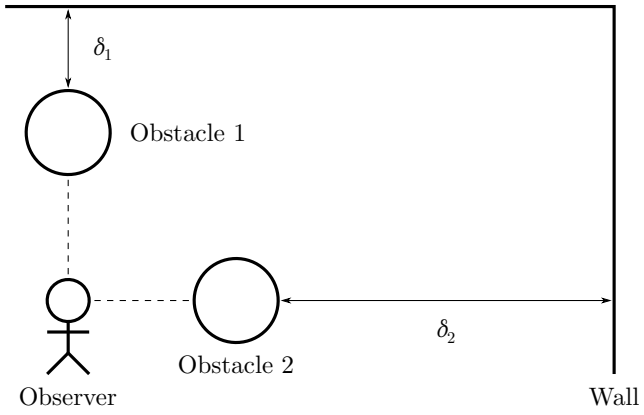
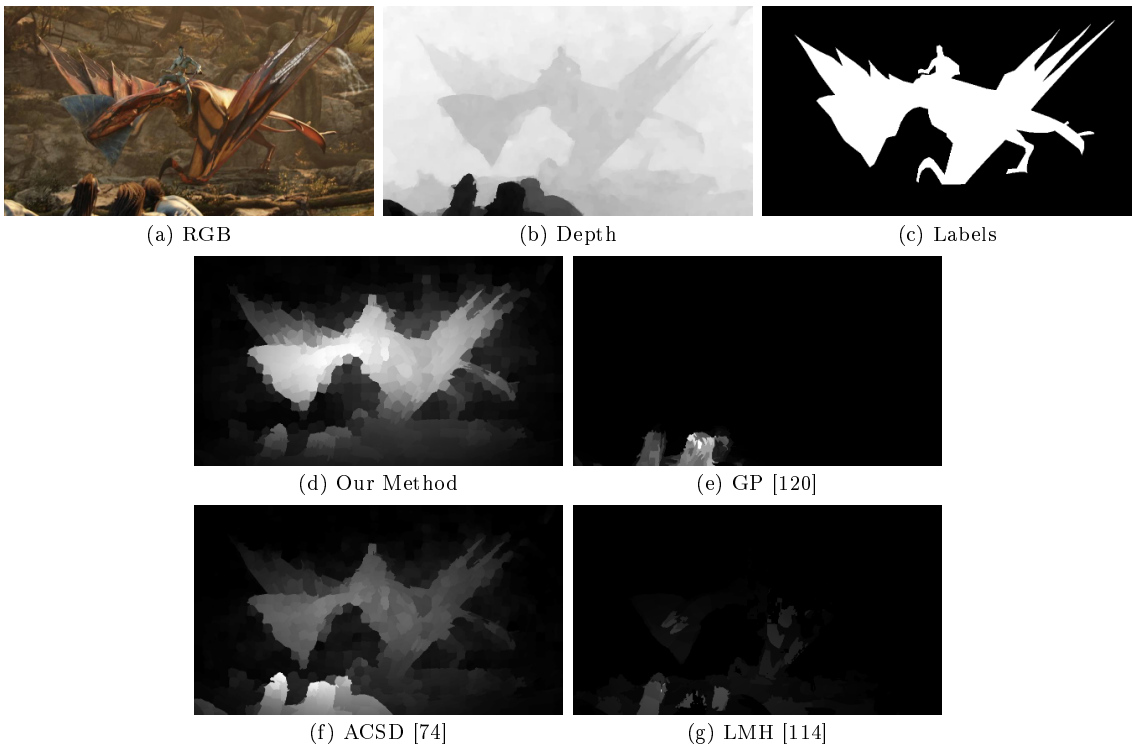


Figure 6-1. Illustration of a problem with applying depth contrast saliency for assistive navigation. Since depth contrast saliency is based on the distance between an object and its surroundings, in this example the depth contrast saliency of Obstacle 1 will be lower than Obstacle 2 from the point of view of the Observer because $\delta_1 < \delta_2$. Thus, a scene representation based on depth contrast saliency would primarily convey Obstacle 2, despite the fact that both obstacles are the same distance from the Observer and pose the same collision risk.

obstacles that pose the same collision risk to the observer are assigned different depth contrast saliency scores. Assigning consistent scores that better reflect structural importance is a crucial step for enabling effective vision processing for navigation.

Aiming to address these issues, we propose a new feature that captures salient structure based on the degree to which an object is bordered by regions of greater depth, called Local Background Enclosure (LBE). We note that salient objects tend to be characterised by being locally in front of surrounding regions, and the distance between an object and the background is not as important as the fact that the background surrounds the object for a large proportion of its boundary. The existence of background in a large spread of angular directions around the object implies pop-out structure and thus high saliency. Conversely, background regions are less likely to exhibit pop-out structure. Thus our proposed depth saliency feature incorporates two components. The first, which is proportional to saliency, is the angular density of background around a region, encoding the idea that a salient object is in front of most of its surroundings. The second feature component, which is inversely proportional



(a) RGB (b) Depth (c) Labels

(d) Our Method (e) GP [120]

(f) ACSD [74] (g) LMH [114]

Figure 6-2. Saliency output on a depth image where foreground depth contrast is relatively low. Our method measures background enclosure of the object to overcome this problem.

to saliency, is the size of the largest angular region containing only foreground, since a large value implies significant foreground structure surrounding the object. This is the first time angular distributions of background directions have been explicitly measured for depth saliency. This feature is shown to be more robust than existing depth contrast-based measures. Further, we validate the proposed depth feature in a saliency system. We demonstrate that our depth feature out-performs state-of-the-art methods from the time of publication when combined with a depth prior, spatial prior, background prior, and Grabcut refinement.

6.2 Local Background Enclosure

In this section we introduce the Local Background Enclosure feature, which quantifies the proportion of the object boundary that is in front of the background. The salient object detection system will be described in Section 6.3. Given an input RGB map I

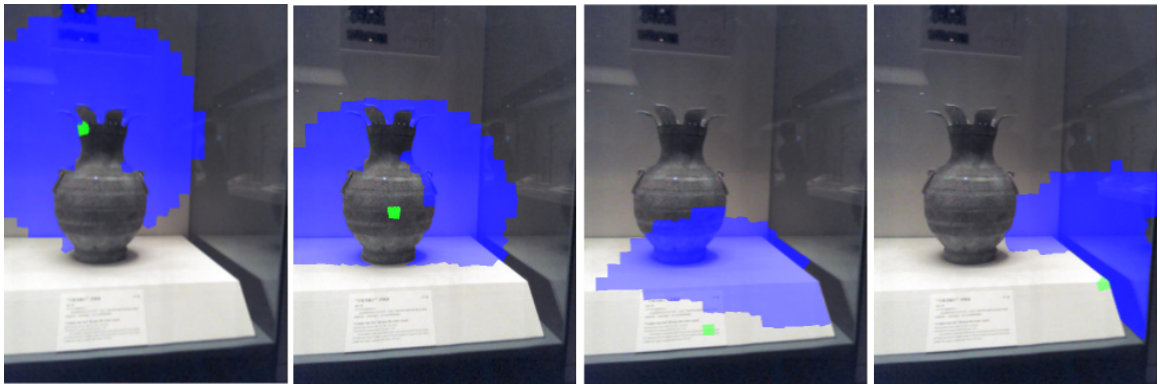


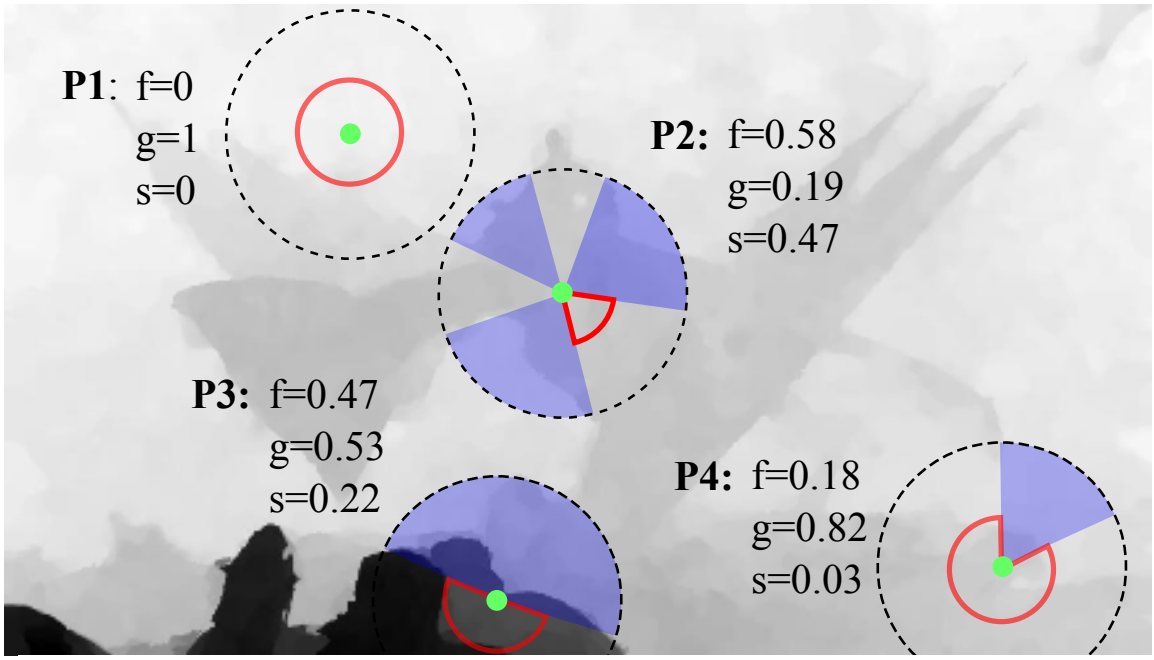
Figure 6-3. Illustration of the local background sets (blue) for four different candidate regions (green). In this example the neighbourhood radius is $r = 200$ pixels, and the depth cutoff is $t = \sigma/2$. Note that patches lying on salient objects tend to be enclosed by the local background set.

and depth map D, we aim to segment the pixels into salient and non-salient pixels. For computational efficiency and to reduce noise from the depth image, instead of directly working on pixels, we oversegment the the image into a set of patches $\mathcal{P}$ according to their RGB value. We denote individual patches as $P \in \mathcal{P}$. We use SLIC [3] to obtain the superpixel segmentation, although our method is flexible to the type of segmentation method used.

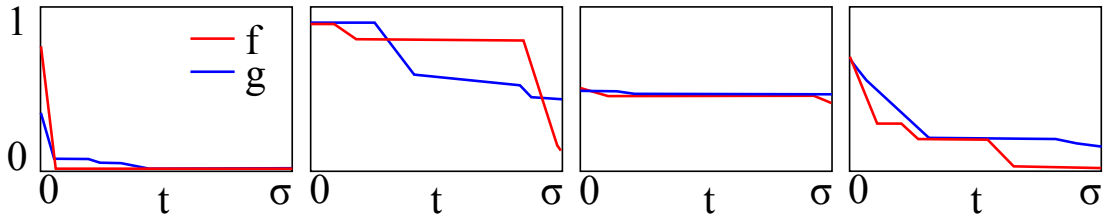
Salient objects tend to be locally in front of their surroundings, and consequently will be mostly enclosed by a region of greater depth, as shown in Figure 6-3. We propose the Local Background Enclosure feature denoted by S_{lbe} based on D. This feature employs an angular density component, F, and an angular gap component, G, to measure the proportion of the object boundary in front of the background.

6.2.1 Angular Density Component

We wish to measure the angular density of the regions surrounding P with greater depth than P, referred to as the local background. We consider a local neighbourhood N_P of P, consisting of all patches within some fixed radius r of P. That is, $N_P = \{Q \mid \|(x_P, y_P) - (x_Q, y_Q)\|_2 < r\}$, where (x_P, y_P) and (x_Q, y_Q) are patch centroids.



(a) Density Functions, t = σ/2



P1: F = 0.05 G = 0.06 S = 0.03 **P2:** F = 0.66 G = 0.78 S = 0.51 **P3:** F = 0.48 G = 0.47 S = 0.23 **P4:** F = 0.22 G = 0.16 S = 0.04

(b) Distribution Functions, t∈[0, σ]

Figure 6-4. Illustration of the background enclosure feature evaluated on the depth image from Figure 1. (a) The density functions computed at image locations marked by the green points with neighbourhood boundaries marked by dotted lines. The blue fill denotes angular regions containing points with greater depth than $t = \sigma/2$ from the center depth, with the maximum gap between these regions marked in red. The values of the angular density component f, the angular gap component g, and saliency $s = f \cdot (1 - g)$ for $t = \sigma/2$ are marked. (b) The distribution functions F, G, and final LBE saliency $S = F \cdot G$ at each point.

We define the local background $\mathcal{B}(P,t)$ of P as the union of all patches within a neighbourhood N_P that have a mean depth above a threshold t from P.

$$\mathcal{B}(P,t) = \bigcup \{Q \in N_P | D(Q) > D(P) + t\}, \tag{6.1}$$

where $D(P)$ and $D(Q)$ denote the mean depth of P and Q respectively.

We define a function $f(P, \mathcal{B}(P,t))$ that computes the normalised ratio of the degree to which $\mathcal{B}(P,t)$ encloses P.

$$f(P, \mathcal{B}(P,t)) = \frac{1}{2\pi} \int_0^{2\pi} \eta(\theta, P, \mathcal{B}(P,t)) \, \mathrm{d}\theta, \tag{6.2}$$

where $\eta(\theta, P, \mathcal{B}(P,t)))$ is an indicator function that equals 1 if the line passing through the centroid of patch P with angle θ intersects $\mathcal{B}(P,t)$, and 0 otherwise. Note that we assume that P has a high compactness [3]. A visualisation of f is shown in Figure 6-4.

Thus $f(P, \mathcal{B}(P,t))$ computes the angular density of the background directions. Note that the threshold t for background is an undetermined function. In order to address this, as frequently used in probability theory, we employ the distribution function, denoted as $F(P)$, instead of the density function f, to give a more robust measure. We define $F(P)$ as:

$$F(P) = \int_0^{\sigma} f(P, \mathcal{B}(P,t)) \, \mathrm{d}t, \tag{6.3}$$

where σ is the standard deviation of the mean patch depths within the local neighbourhood of P. This is given by $\sigma^2 = \frac{1}{|\mathcal{B}(P,0)|} \sum_{Q \in \mathcal{B}(P,0)} \left(D(Q) - \overline{D}\right)^2$, where $\overline{D} = \frac{1}{|\mathcal{B}(P,0)|} \sum_{Q \in \mathcal{B}(P,0)} D(Q)$. This implicitly incorporates information about the distribution of depth differences between P and its local background.

6.2.2 Angular Gap Component

In addition to the angular density $F(P)$, we introduce the angular gap statistic $G(P)$. As shown in Figure 6-4, even though P_2 and P_3 have similar angular densities, we

would expect P_2 to have a significantly higher saliency since the background directions are more spread out. To capture this structure, we define the function $g(P, Q)$ to find the largest angular gap of Q around P and incorporate this into the saliency score.

$$g(P, Q) = \frac{1}{2\pi} \cdot \max_{(\theta_1, \theta_2) \in \Theta} \{|\theta_1 - \theta_2|\}, \tag{6.4}$$

where Θ denotes the set of boundaries (θ_1, θ_2) of angular regions that do not contain background:

$$\Theta = \{(\theta_1, \theta_2) \mid \eta(\theta, P, Q) = 0 \ \ \forall \theta \in [\theta_1, \theta_2]\}. \tag{6.5}$$

A visualisation of g is shown in Figure 6-4.

We define the angular gap statistic as the distribution function of $1 - g$:

$$G(P) = \int_0^{\sigma} 1 - g(P, \mathcal{B}(P, t)) \, \mathrm{d}t. \tag{6.6}$$

The low level Local Background Enclosure value is thus given by:

$$S_{\mathrm{lbe}}(P) = F(P) \cdot G(P). \tag{6.7}$$

Figure 5-8 shows the generated saliency map on some example images. Note that the pop-out structure corresponding to salient objects is correctly identified. Depth contrast features fail to detect the objects, or exhibit high false positives.

6.3 Saliency Detection System

We construct a system for salient object detection using the proposed feature. Specifically, we reweight the Local Background Enclosure feature saliency using depth and spatial priors, and then refine the result using Grabcut segmentation. An overview of our system is given in Figure 6-5.

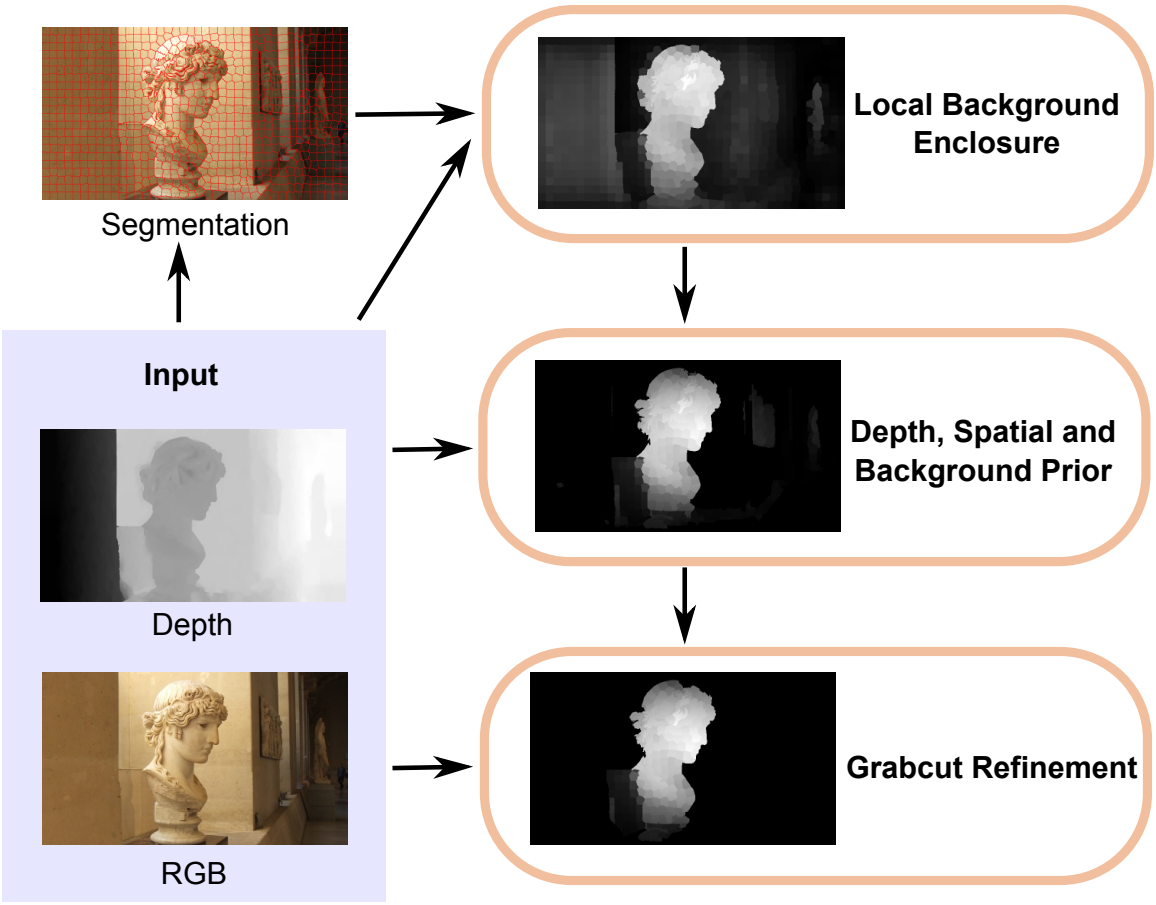


Figure 6-5. Overview of our saliency detection system. Given an RGB-D image and superpixel segmentation, we first compute our Local Background Enclosure feature, then apply depth, spatial, and background priors, and finally refine the result using Grabcut segmentation.

6.3.1 Depth, Spatial, and Background Prior

We apply three standard saliency priors to reweight the low level saliency map S_{lbe}, producing the prior adjusted saliency map S_{prior}. First, we perform absolute depth reweighting using a depth prior to modulate the saliency of pixels with depth greater than the median depth of the image, as in [74]. Following this, we apply a spatial prior that re-weights patch saliency based on distance from the image center [99]. Finally, we use the background prior map described in [164] to reweight saliency based on a measure of boundary connectedness. Prior application is a standard step in existing saliency systems, further details are available in Chapter 2.

6.3.2 Grabcut Segmentation

The saliency map S_{prior} may contain inaccurate foreground boundaries for parts of the object that do not exhibit strong pop-out structure. Boundary refinement is a common post-processing step employed in existing salient object detection systems (*e.g.* [26, 114, 106, 120, 58]). We use the same graph cut based boundary refinement as in Chapter 5 to improve object boundaries using appearance information. The graph cut foreground model is initialized with a binary mask obtained by applying a threshold of 0.8 to S_{prior}. The output Grabcut segmentation mask M_{gc} is used to prune non-foreground areas from S_{prior}. The refined saliency map is thus given by

$$\mathcal{S}_{\text{lbe}}(x, y) = M_{\text{gc}}(x, y) \cdot S_{\text{prior}}(x, y). \tag{6.8}$$

6.3.3 Implementation Details

The discrete version of the angular density function f is implemented using a histogram-based approximation, denoted as $\tilde{f}$. Let $h\left(i, P, \mathcal{B}\left(P, t\right)\right)$ be an n bin polar occupancy histogram, where bin i is 1 if the corresponding angular range contains an angle between the centroids of P and a patch in $\mathcal{B}\left(P, t\right)$, and 0 otherwise. We set $\tilde{f}$ to be equal to the fill ratio of h.

$$\tilde{f} = \frac{1}{n} \sum_{i=1}^{n} h\left(i, P, \mathcal{B}\left(P, t\right)\right). \tag{6.9}$$

The distribution function F is computed numerically using $\tilde{F}$ by sampling $\tilde{f}$ at m equally spaced points across the integration range such that:

$$F(P) = \frac{1}{m} \sum_{i=1}^{m} \tilde{f}\left(P, \mathcal{B}\left(P, \frac{i \cdot \sigma}{m}\right)\right). \tag{6.10}$$

Similarly, we define $\tilde{G}$ to evaluate G:

$$\tilde{G}(P) = \frac{1}{m} \sum_{i=1}^{m} \left(1 - \frac{1}{2\pi} \cdot g\left(P, \frac{i \cdot \sigma}{m}\right)\right). \tag{6.11}$$

6.4 Experiments

The performance of our saliency system is evaluated on two datasets for RGB-D salient object detection. RGBD1000 [114] contains 1000 RGB and structured light depth images. NJUDS2000 [74] contains 2000 RGB and disparity images computed from stereo image pairs.

The proposed Local Background Enclosure feature is compared against the following state-of-the art contrast-based depth features: multi-scale depth-contrast (LMH-D) [114]; global depth contrast (GP-D) [120]; and ACSD [13]. We also include versions of LMH-D and GP-D with signed depth, denoted LMH-SD and GP-SD respectively, where neighbouring patches with a lower average depth do not contribute to the contrast measure of a patch. Additionally, in order to verify the contribution of using the distribution functions, we compute the product of the density functions $f(P,t) \cdot g(P,t)$ with fixed threshold $t = \sigma/2$.

We then evaluate the contribution of prior application and Grabcut refinement on our salient object detection system on both datasets. Finally, we compare our salient object detection system with three state-of-the-art RGB-D salient object detection systems: LMH [114], ACSD [74], and a recently proposed method that exploits global priors, which we refer to as GP [120]. We also include comparisons with the state-of-the-art 2D saliency algorithms DRFI [71] and DSR [89], which were found to be top ranking methods by a recent study [20].

6.4.1 Evaluation Metrics

We present the precision-recall curve and mean F-score to evaluate algorithm performance. The F-score is computed from the saliency output using an adaptive threshold equal to twice the mean of the image [1]. Note that the F-score is calculated as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{6.12}$$

where $\beta = 0.3$ to weigh precision more than recall [1].

6.4.2 Experimental Setup

We set $n = 32$ histogram bins and $m = 10$ evaluation steps in our implementation of F and G respectively. These two values were found to provide a good trade-off between accuracy and efficiency for general use. The radius of the neighbourhood N_P should be set to equal the expected radius of the largest object to detect, thus we set it to half the image diagonal for general use. We use SLIC [3] on the colour image to generate the set of patches, with the number of patches set to the length of the diagonal of the image in pixels.

6.5 Results and Discussion

This section will first present the overall results of our saliency detection system. Following this, it will then provide an analysis of the improvements offered by our low-level LBE feature over depth contrast-based features used in existing systems. Failure cases will then be discussed, and finally output from each stage of our saliency detection system will be presented.

6.5.1 Saliency Detection System Results

Figure 6-7 shows that our saliency system outperforms all other state-of-the-art RGB-D salient object detection systems. Our saliency system achieves the highest F-score on both datasets, with GP obtaining the second best performance. In addition to the highest F-score, our method exhibits the highest recall among the depth-based methods on both datasets, reflecting the fact that our depth feature correctly identifies a greater portion of the foreground compared to contrast-based methods. From Figure 6-7a we see that our method has the highest PR curve on RGBD1000. Figure 6-7b shows that our system has high precision up to around 0.65 recall, with superior performance in the region of high precision. This demonstrates that our feature is able to identify salient structure from depth more effectively than existing contrast-based methods. With the exception of DRFI on RGBD1000, the RGB methods

perform worse than most depth-aware methods.

Figure 5-8 shows the output of our salient detection system compared with state-of-the-art methods. Note that the other methods tend to have a high number of false positives due to depth contrast in background regions, for example depth change across a flat table is registered as salient by ACSD in the second row. The angular statistics employed by our depth feature provide a more robust measure of salient structure.

6.5.2 Comparison with Contrast-based Depth Features

The LBE feature outperforms the contrast-based depth features used in state-of-the-art systems (Figures 6-6a and 6-6b). The performance of the depth features of GP and LMH is significantly improved when excluding patches with lower depth than the candidate patch during contrast computation. It can also be seen that using the distribution function gives improved results compared to using the density functions evaluated at a fixed threshold t.

We now provide a detailed analysis on the nature of the improvements gained by our depth feature, specifically the ways in which false negatives and false positives produced by contrast-based methods are reduced when using LBE. Note that since we are comparing depth features, no colour information is used at this stage. Additionally, no priors are applied to the generated saliency maps.

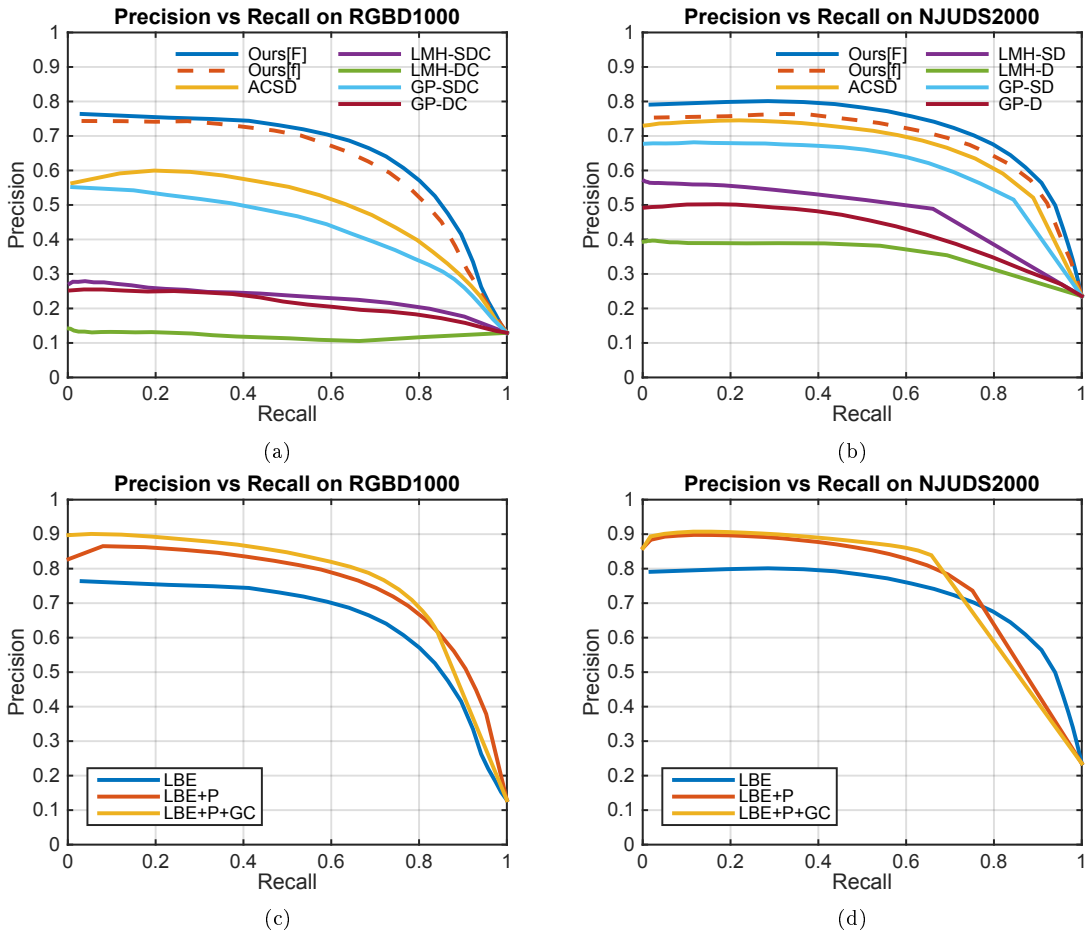


Figure 6-6. PR curves showing performance of the LBE feature against contrast-based depth features on (a) RGBD1000 and (b) NJUDS2000. Ours[f] refers to LBE computed with the angular fill and gap density functions f and g, and Ours[F] refers to LBE computed with the distribution functions F and G. PR curves showing the effect of each component of the saliency system on (c) RGBD1000 and (d) NJUDS2000. LBE denotes our proposed local background enclosure feature, and P and GC refer to prior application and Grabcut refinement respectively.

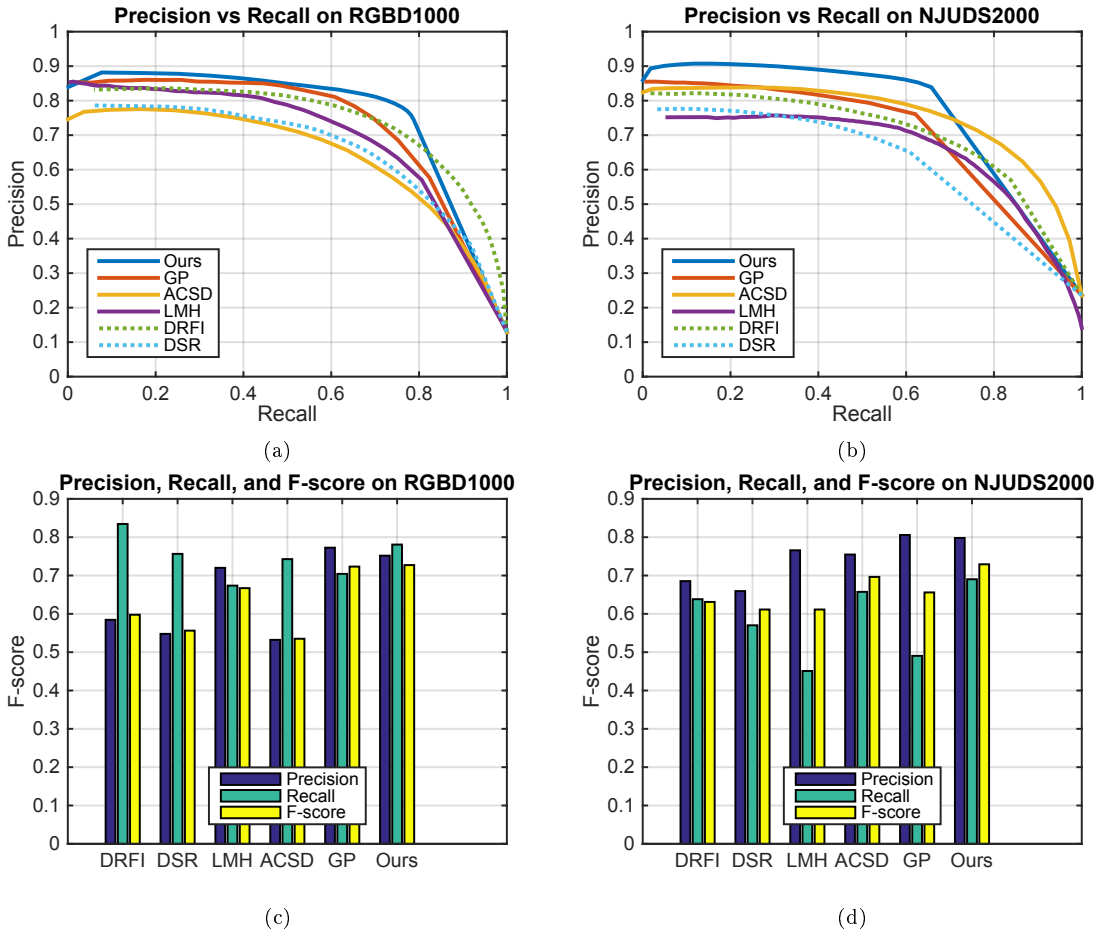
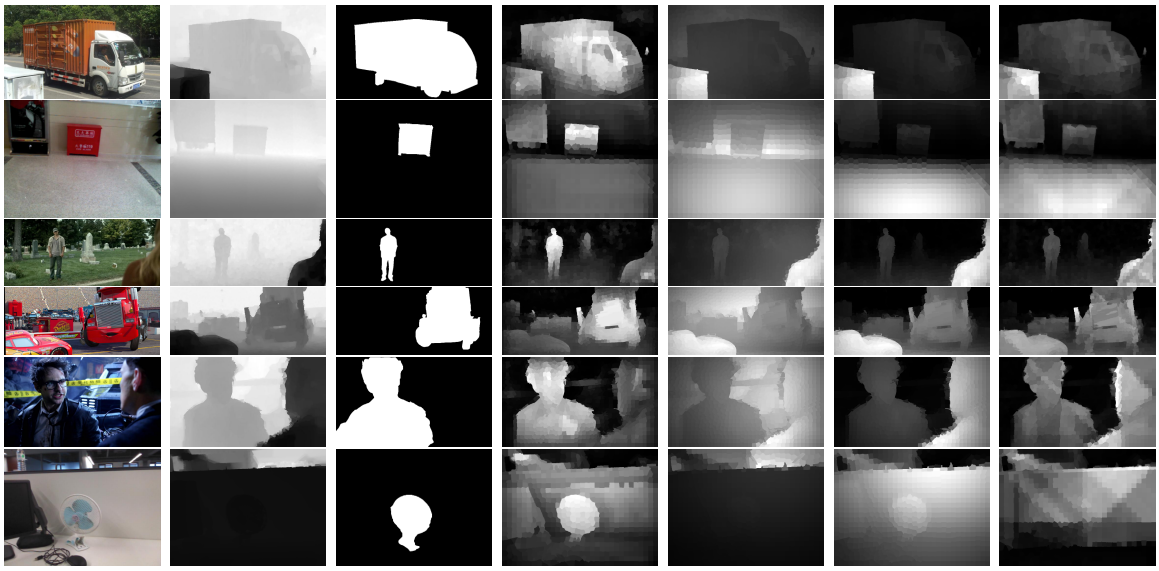


Figure 6-7. A comparison of our proposed saliency system against the state-of-the-art RGB-D saliency systems DRFI [71], DSR [89], LMH [114], ACSD [74], and GP [120]. The PR curve of each method is shown on (a) RGBD1000 and (b) NJUDS2000. The precision, recall, and F-measure of each method is shown on (c) RGBD1000 and (d) NJUDS2000.

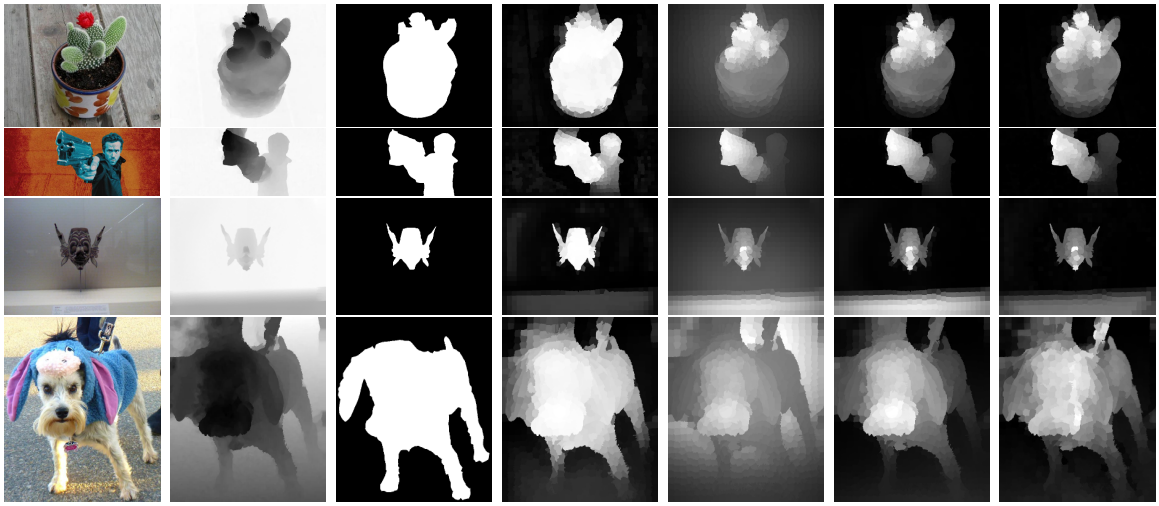


(a) Colour (b) Depth (c) Labels (d) LBE (Ours) (e) DC [120] (f) SDC [120] (g) ACSD [74]

Figure 6-8. Examples of performance on images where the foreground has low depth contrast compared to the background using the raw depth saliency features of LBE, state-of-the-art methods ACSD[74], DC[120], and SDC[120] which is a signed variant of depth contrast. The depth contrast based methods perform poorly, while our method identifies the salient objects.

6.5.2.1 Reducing False Negatives: Low Contrast Foreground

A common pitfall of existing contrast-based depth features is sensitivity to depth difference magnitude. These features produce false negatives when the object has lower contrast than the background. Figure 6-8 shows example scenes where depth contrast based methods incorrectly assign a low saliency score to the salient object because it has relatively low depth contrast. For example in the first row, the white box in the bottom right corner of the image has the greatest depth difference with the surroundings. This object is not salient according to the ground truth, however existing methods identify the box as the salient object. In these cases our depth feature correctly identifies the salient object based on its pop-out structure as measured using local background enclosure.

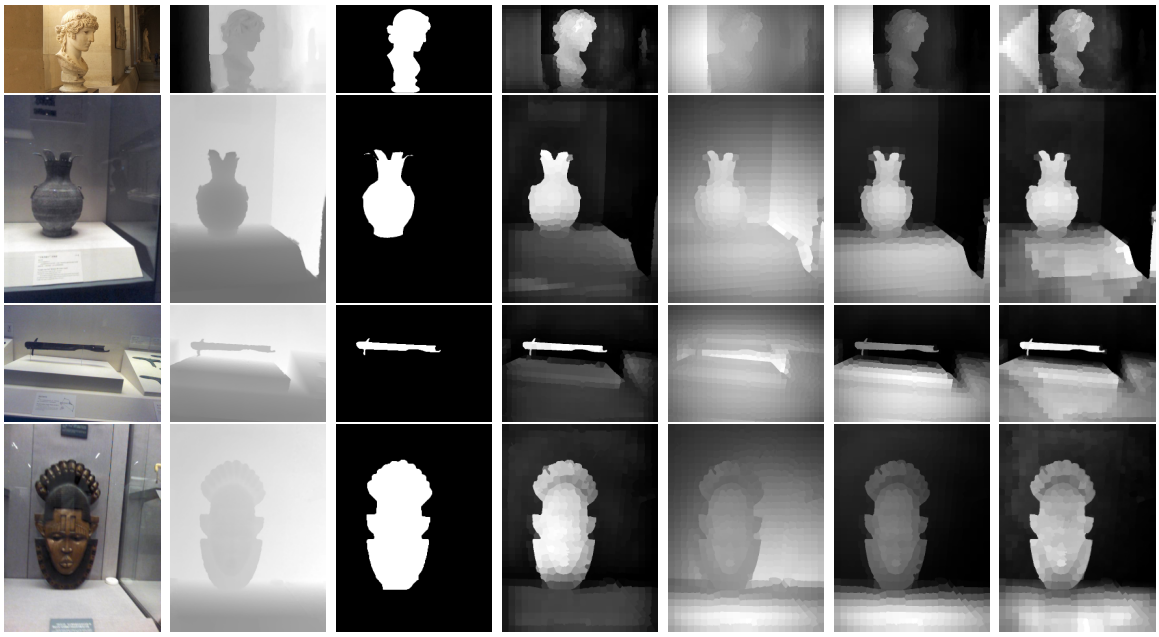


(a) Colour (b) Depth (c) Labels (d) LBE (Ours) (e) DC [120] (f) SDC [120] (g) ACSD [74]

Figure 6-9. Examples of objects containing different depth contrast values, producing a non-uniform saliency response from the depth contrast based methods ACSD [74], DC [120] and SDC [120] which is a signed version of depth contrast. Our method LBE is able to obtain a uniform saliency response across an object when the object pops out from the surroundings.

6.5.2.2 Reducing False Negatives: Objects with Large Depth Range

Salient objects that contain a relatively large range of depth values tend to have a high variation in the saliency response across the object, as shown in Figure 6-9. For example, in the first row, there is a significant difference between the saliency values of the top of the plant and the pot for contrast based features. For these images and others like them, our feature produces a more uniform response across objects which have a pop-out shape.

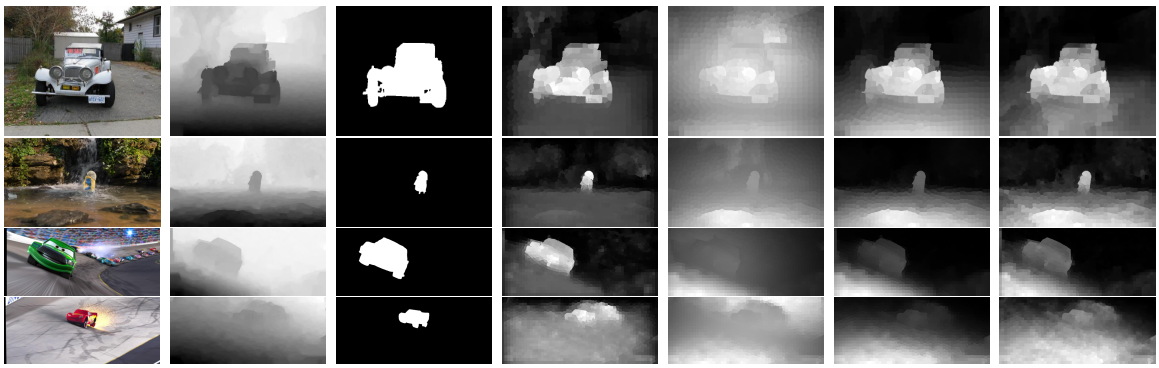


(a) Colour (b) Depth (c) Labels (d) LBE (Ours) (e) DC [120] (f) SDC [120] (g) ACSD [74]

Figure 6-10. Examples of performance on images where the background exhibits high depth contrast using the raw depth saliency features of LBE, state-of-the-art methods ACSD[74], DC [120], and SDC [120] which is a signed variant of depth contrast. In this type of situation using LBE to measure pop-out structure more robustly identifies foreground regions compared to measuring depth contrast.

6.5.2.3 Reducing False Positives: High Contrast Background

Background structure adjacent to a large depth drop-off is a common source of false positives for depth contrast methods, producing high depth contrast values in the background region, as shown in Figure 6-10. For example, in the first row, the wall on the left is assigned a relatively high saliency by depth contrast based features because it has large depth difference with the adjacent region. Since background structure usually does not have a pop-out shape, our feature is able to suppress these regions.

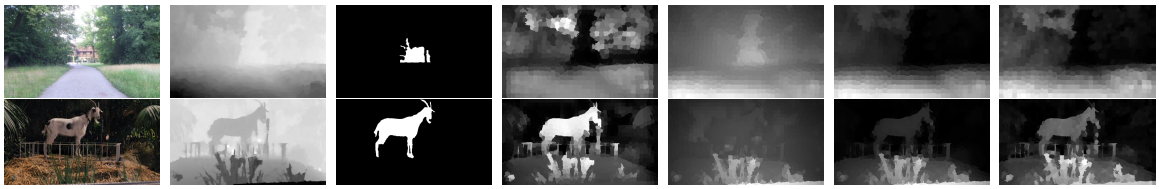


(a) Colour (b) Depth (c) Labels (d) LBE (Ours) (e) DC [120] (f) SDC [120] (g) ACSD [74]

Figure 6-11. Examples of saliency output on scenes with large angled planar surfaces, which have a high depth contrast, using the raw depth saliency features LBE, ACSD[74], DC [120], and SDC [120] which is a signed variant of depth contrast. Planar surfaces are generally unavoidable, and produce a high saliency response for depth contrast based methods. These types of surfaces do not exhibit pop-out structure however, and are therefore assigned a low saliency score by our depth feature.

6.5.2.4 Reducing False Positives: Angled Planar Surfaces

Flat surfaces that are angled towards the camera are one particularly common type of background structure that exhibits depth contrast. These surfaces frequently produce false positives for depth contrast methods, since points along the surface can have a wide range of depth values. Some examples are shown in Figure 6-11. Our method significantly reduces the false positives caused by this type of structure, since depth difference across the surface is ignored, and since large planar surfaces tend to have a low background enclosure.



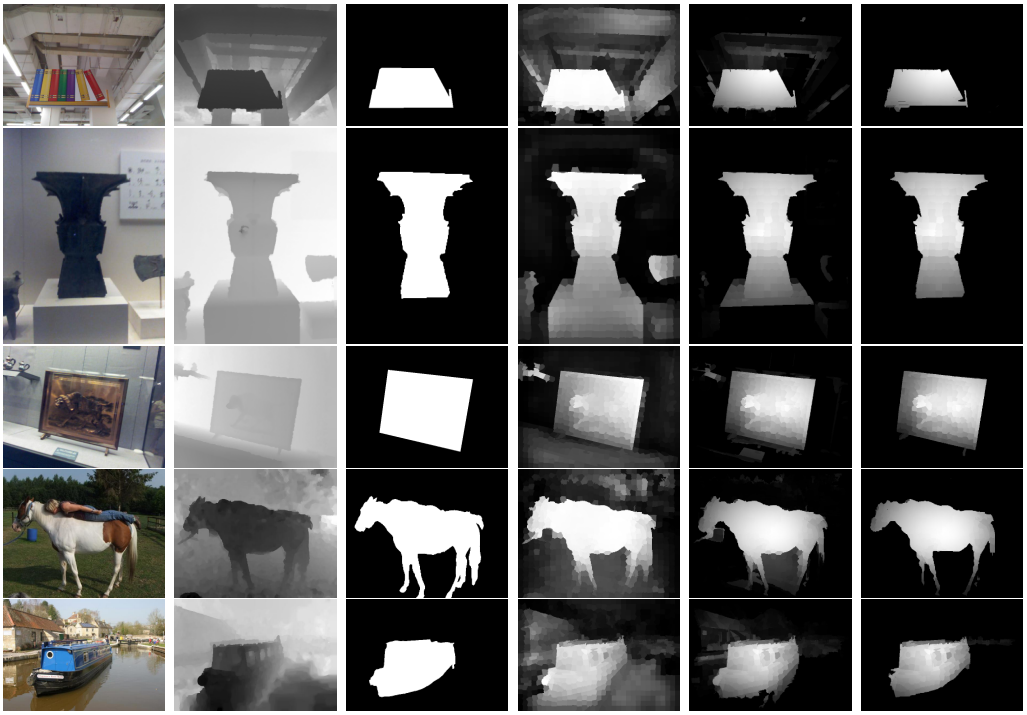
(a) Colour (b) Depth (c) Labels (d) LBE (Ours) (e) DC [120] (f) SDC [120] (g) ACSD [74]

Figure 6-12. Examples of failure cases, showing saliency output from the raw depth saliency features LBE, ACSD [114], DC [120], and SDC [120], which is a signed version of depth contrast. In the first row, the object is surrounded in all directions by closer surfaces. This is a rare occurrence, as salient objects tend to be in front of their surroundings. The second row shows a situation where a background region has strong pop-out structure. This leads to false positives for all methods, and our method produces the best result in this case.

6.5.3 Failure Cases

Since our method measures pop-out structure, it does not produce good results when the salient object is surrounded in all directions by background with lower depth. An example is shown in Figure 6-12. Note that this is a rare occurrence, and the other depth saliency methods with the exception of DC also produce poor results in this case. In these situations, it is questionable whether the object can be considered to be salient. Note that DC produces the best results in this image because it does not assume that salient objects are in front of the background, however this leads to poor performance on the datasets.

Occasionally the background can have some degree of pop-out structure, such as the grass in the second row of Figure 6-12, leading to false positives from our feature. However the response is generally weaker than for the salient object, and it is a less common occurrence than the background having high depth contrast. Our depth feature still produces the best overall result compared to contrast based depth features, which are also affected by this problem.



(a) RGB (b) Depth (c) Labels (d) LBE (e) LBE+P (f) LBE+P+G

Figure 6-13. Output of the different stages of our salient object detection system. LBE denotes our proposed depth feature, LBE+P shows the result of depth, spatial, and background prior application, and LBE+P+G illustrates the final output of our salient object detection system after applying Grabcut refinement.

6.5.4 Saliency Detection System: LBE, Priors, and Grabcut Outputs

Figures 6-6c and 6-6d show the quantitative contributions of each of the three stages of our saliency detection system. We will now present examples showing the output from each stage of our system in Figure 6-13. First the LBE feature is applied to the depth image, identifying the salient object and sometimes producing a non-zero response for background regions with pop-out structure. These background regions are trimmed based on depth, spatial position and colour during the prior application stage. The resulting map is further pruned in the Grabcut refinement stage.

6.6 Chapter Summary

In this chapter, we have proposed a novel depth feature that exploits local depth background enclosure to detect salient objects in RGB-D images. While existing work in structural salient object detection employs depth contrast as the main feature, our proposed LBE feature captures the spread of angular directions that are background with respect to the candidate region and the object that it is part of. This addresses a fundamental issue with applying depth contrast saliency for prosthetic vision scene representation, mitigating the effect of high contrast background regions on saliency detection. Our approach also implements the intuition that depth saliency should not be dependent on the exact distances between objects, but rather the general arrangement of structure in the scene.

We have shown through our results that the LBE feature is a more useful structural cue for salient object detection than depth contrast features used in existing methods. Specifically, the raw LBE feature improves saliency detection accuracy compared to depth contrast features across a wide variety of scenes containing low contrast foreground, objects with a large range of depth values, high contrast background, and angled planar surfaces. Furthermore, refining the low level LBE saliency with priors and Grabcut refinement produces better results than previous state-of-the-art salient object detection systems on two publicly available RGB-D salient object detection datasets. This demonstrates that our LBE feature is able to identify salient structure from depth more effectively than existing contrast-based methods.

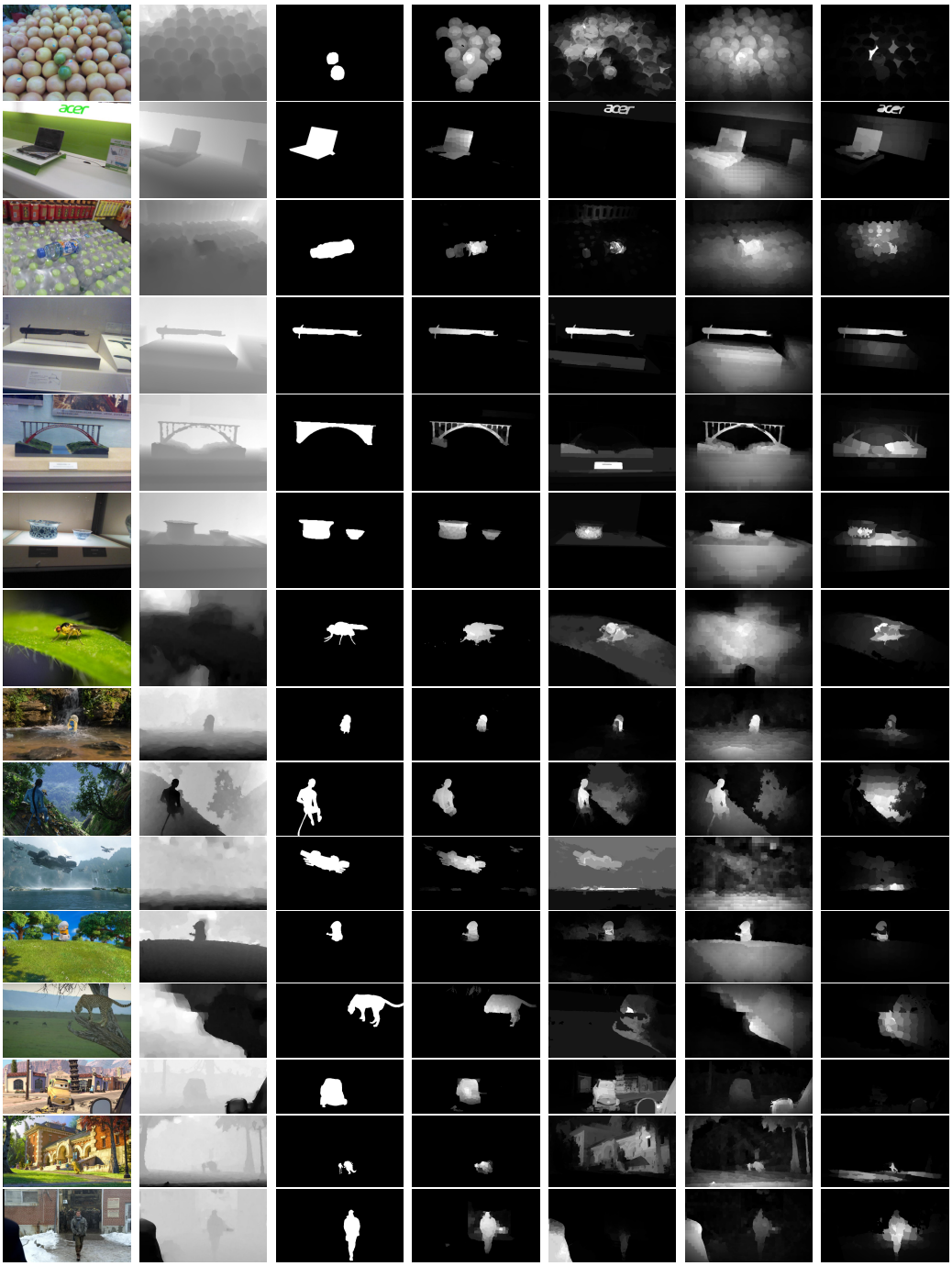
The findings presented in this chapter provide novel insight into what kind of object structure is salient to the human visual system, addressing fundamental limitations of existing methods and offering a robust basis for future prosthetic vision scene representations on object-related tasks.

6.7 Summary of Technical Chapters

Chapter 6 concludes the technical portion of the thesis. We have presented four chapters that comprehensively address the thesis subproblems of identifying salient edges for conveying scene shape, and identifying salient objects for approximating biological visual attention direction.

Chapters 3 and 4 have investigated the extraction of salient edges that capture scene shape by detecting surfaces with irregular shape compared to their surroundings. Chapter 4 has improved the recovery of relevant structure by incorporating high level information to refine the detection of salient edges corresponding to boundaries of interest to humans in the scene. The findings from these chapters can be used to develop visual representations that convey scene shape for performing physical tasks such as navigation with prosthetic vision.

Following this, Chapter 5 has investigated the extraction of general object structure that is salient to the human visual system through analysis of surface shape. Finally, in Chapter 6 we have proposed a new model of structural saliency that addresses fundamental limitations of previous methods, and provides further insight into what kind of object structure is salient to the human visual system. The work in these chapters can be used to identify structure that is likely to be important when first viewing a scene, and enables vision processing methods to more closely emulate biological attention deployment mechanisms. In particular, this work could be applied to develop scene representations for obstacle avoidance, landmark-based orientation, and tabletop or grasping tasks.



(a) RGB (b) Depth (c) Labels (d) Ours (e) GP [120] (f) ACSD [74] (g) LMH [114]

Figure 6-14. Comparison of output saliency maps produced by our salient object detection system against the output of GP [120], ACSD [74], and LMH [114]. Our robust LBE depth feature allows for a more accurate final saliency map compared to methods using contrast based depth features.

Chapter 7

User Study Evaluation: Surface Irregularities

The preceding four technical chapters have proposed structural saliency techniques for scene representation with prosthetic vision, addressing the thesis subproblems of detecting salient edges and salient objects based on scene structure. We now aim to determine whether the enhanced presentation of scene structure offered by our proposed methods leads to measurable improvements in user performance during practical use of a prosthetic vision display. In order to achieve this, Chapters 7 and 8 will present evaluations of structure-based vision processing methods on navigation tasks.

In this chapter, we evaluate the surface irregularities method proposed in Chapter 3 for navigation with prosthetic vision, through a pilot user study in which normally sighted participants complete a navigation task using SPV. The aim of the study is to determine whether surface irregularities offers improvement over standard methods, as measured by the ability to facilitate safe navigation by reducing collisions during navigation. This is evaluated in order to verify the suitability of surface irregularities for further testing in a clinical trial. The same experiment was subsequently repeated in a separate clinical trial with retinal implant users [13]. Results from these clinical trials are discussed, but were not published as part of this thesis.

Note that surface irregularities is the only proposed method tested in our pilot

study. This study was run as a evaluation of general scene structural representations, for which the salient object detection methods were deemed less suitable. Additionally, the improved edge detection system from Chapter 4 had not been completed at the time the study was undertaken.

This chapter is organised as follows. First, an introduction to the study is given in Section 7.1 and an overview of the visual representations tested in the study is given in Section 7.2. The experiment design, including details of the task, environment, obstacles, metrics and experimental procedure, is outlined in Section 7.3. Section 7.4 presents the results of the pilot study, and Section 7.5 provides discussion, including discussion of the results from the subsequent clinical trial [13] using surface irregularities. Section 7.6 concludes the chapter.

7.1 Introduction

Performing navigation with a visual prosthesis is challenging because of the low resolution and dynamic range of the prosthetic vision display [39]. These display constraints make it difficult to interpret scene structure and easy to miss important environment details such as small or low-contrast obstructions [104]. The proposed surface irregularities scene representation in Chapter 3 aims to address this problem by detecting and conveying structurally salient scene components that are important for navigation. In this chapter, we evaluate the effectiveness of the surface irregularities visual representation through a user study, in which participants use the visual representation to perform the task of moving through a corridor-like environment containing obstacles, while avoiding collisions with the obstacles and environment walls. This is intended to serve as a pilot study, involving normally-sighted participants performing the task using SPV, for validating the feasibility of the surface irregularities visual representation before further testing in a clinical trial with implant users. We also present a discussion of the results from the subsequent clinical trial [13].

Evaluating the effectiveness of vision processing methods for navigation is challenging, due to the high likelihood for confounding factors and the potential complex-

ity of the task [150]. The key difficulties of evaluating navigation performance in our study, and our approach for addressing each difficulty, are as follows:

- The performance measures for successful navigation with prosthetic vision are not as obvious as performance measures for other prosthetic vision tasks such as reading or facial recognition. While there is a large field of work on assessing orientation and mobility performance for specific applications such as measuring progressive incapacity with ageing [145], exploring the effect of various visual factors on mobility [100], and assessing mobility training outcomes [137], the act of quantifying general navigation performance requires reducing the entire sequence of participant actions on a complex task to a relatively small set of performance measures, and some information will invariably be lost. We therefore narrow the scope of the problem by measuring the ability of the visual representation to enable *safe* navigation, due to the emphasis on safety when identifying desired outcomes for mobility [23, 39]. Thus, we define navigation as the process of performing a forward ambulatory motion through traversable space while avoiding obstructions, and we primarily quantify navigation performance through the number of collisions with the environment. This is a standard metric used in previous studies [111, 104]. We also record whether the participant is correctly oriented at the conclusion of each trial.

- There are many environmental factors that affect navigation performance, including lighting conditions, scene colouring, and environment shape and arrangement [44]. We mitigate these factors by using a custom-built environment in which navigation conditions can be reproduced consistently. In particular, scene lighting and contrast are controlled, and the expected difficulty of navigation due to obstacle placement is kept consistent across trials.

- Performance is also affected by factors that are particular to an individual user, such as individual task-related strategies and preferences, familiarity with the task, and physical and mental condition [44]. We aim to reduce these effects by providing training to each participant, ensuring that all participants have:

the same understanding of the task, visual representations, and basic navigation strategies; and, adequate experience with the task before beginning the trial. Additionally, frequent breaks are provided in between trials to ensure participant comfort and reduce fatigue.

- The use of SPV introduces approximations to the phosphene display that may impact performance compared to real implant use. However, previous literature and our work indicates that there are enough similarities in the methodology that the responses will provide valid data that can act to inform the clinical trials. Therefore, the experiment is run as a comparative study, in which our surface irregularities vision processing method is compared against standard baseline methods, since the SPV-induced approximations affect all visual representations to a similar degree. This provides information on the relative performance of the different methods, and thus determines whether surface irregularities leads to an improvement in task performance.

We will now detail the different vision processing methods tested in this study, followed by further details on the navigation task and experiment environment.

7.2 Visual Representations

In this study, the surface irregularities scene representation is compared against two baseline methods: the standard intensity representation, and the Augmented Depth structural representation. Details of each visual representation are given below.

1. **Surface Irregularities**: This is our proposed scene representation, which displays scene structure by highlighting regions with locally contrasting surface shape as a low level representation of the environment. The brightness function of the surface irregularities visual representation is given in Equation 3.6. See Chapter 3 for more details on generating the surface irregularities map.

2. **Standard Intensity**: The standard downsampled intensity scene representation. This representation conveys scene appearance, sampling the filtered

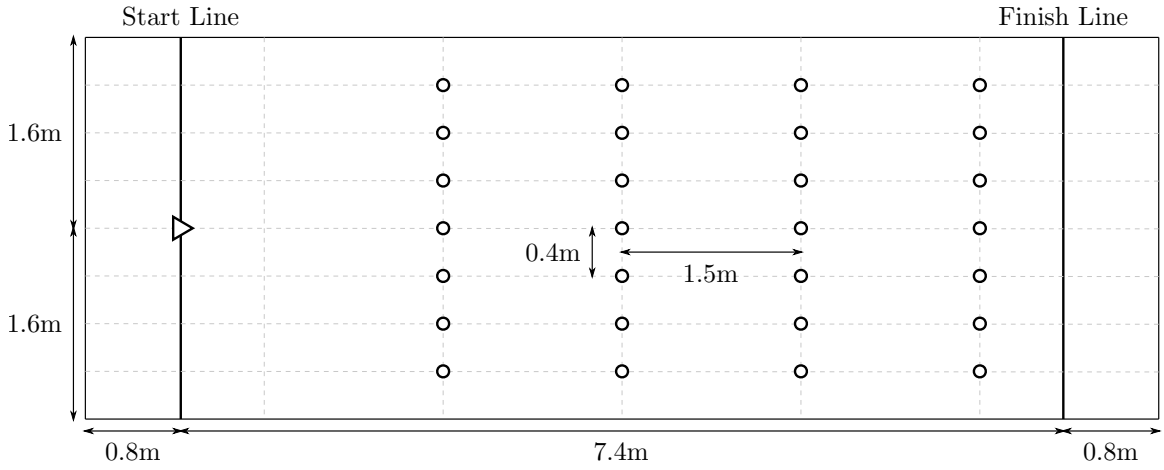


Figure 7-1. A top-down diagram of the experimental environment. The participant start position is marked as a triangle, and possible obstacle locations are denoted with circles.

grayscale camera image at the projected phosphene points and rendering the resulting phosphene image. We use the phosphene brightness function given in Equation 2.14. Comparison with this representation quantifies the improvement from conveying structurally salient regions rather than scene appearance.

3. **Augmented Depth**: An existing structural representation for navigation, which aims to convey obstacle locations to the user. Phosphene brightness is suppressed on the ground plane, increasing contrast between the ground and objects resting on the ground. The brightness function of the Augmented Depth visual representation is given in Equation 2.16. Comparison between our surface irregularities method and Augmented Depth provides an indication of the benefit from conveying general scene structure in addition to obstacle locations.

7.3 Experiment Design

This study aims to compare the effectiveness of the surface irregularities, standard intensity, and Augmented Depth scene representations for navigation, that is, performing a forward ambulatory motion through traversable space while avoiding obstructions. Specifically, performance of each visual representation is measured on the

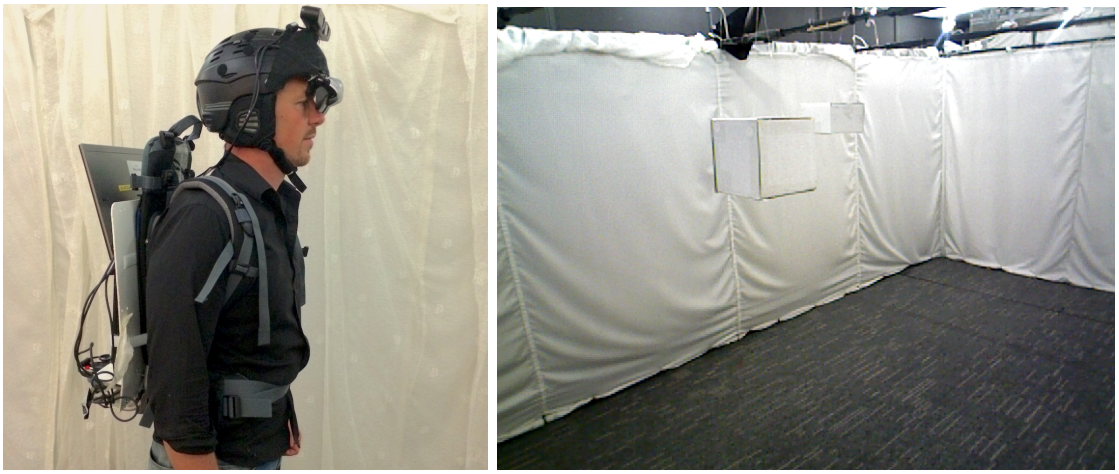
task of moving from one end of a custom built corridor-like environment to the other while avoiding low contrast obstacles positioned within the environment. This task tests whether the participant is able to use the scene representation to detect and avoid contact with the obstacles, and to infer their position and orientation within the environment in order to maintain an appropriate direction of travel.

The study uses a randomised controlled design with repeated measures to evaluate the performance of all visual representations for each participant. For each participant, the experiment consists of a number of trials, *i.e.* traversals through the environment with a given scene representation. The order of presentation of the visual representations is counter-balanced, controlled, and randomly allocated by a computerised system, in order to improve the validity of the trial and limit confounding factors such as fatigue and learning effects. The number, size, and placement of obstacles is also controlled and randomly allocated by a computerised system, while ensuring a sufficiently challenging setup for each trial. Further details of the obstacle randomisation process are available in Section 7.3.3.

Two participants took part in the pilot study. Participant recruitment for the study was performed via email correspondence. Both participants were male, between the ages of 18 and 46, and had normal or corrected-to-normal vision of at least 20/20. Each participant aimed to accomplish 10 trials for each visual representation. The ethical component of this research was approved by the Australian National University Human Research and Ethics Committee.

7.3.1 Mobile SPV System

Participants performed the trials using a mobile SPV system, shown in Figure 7-2a, which provided a real-time rendering of the user's surroundings as they moved through the environment. The SPV system consisted of a head mounted RGB-D camera, head mounted display, and a backpack mounted laptop. The RGB-D camera captured a view of the environment, which was sent to the laptop. The laptop then performed vision processing on this camera image, and rendered the phosphene visualisation. The simulation display was composed of a rectangular 5×4 (width$\times$height)



(a) (b)

Figure 7-2. (a) The mobile prosthetic vision simulation system, which consists of a head mounted RGB-D camera, head mounted display, and a processing unit worn on the back. Note that participants additionally wore a head shroud, not shown here, to block incoming light vision. (b) A view of the experimental environment containing two low contrast overhanging obstacles.

grid of 20 phosphenes rendered within a 15×12 degree field of view. Phosphenes were rendered as circular Gaussian spots with 8 discrete output levels, where phosphene brightnesses between two output are mapped to the closest brightness level. We use the rendering software implementation described in [104]. The phosphene visualization was synchronously presented centrally to both eyes of the participant through an 800×600 head mounted display. A dark cloth shroud fitted over the helmet blocked all incoming light, including the participant's peripheral vision outside the head mounted display.

7.3.2 Navigation Environment

An artificial navigation environment was constructed for the user study in order to facilitate control of random variables and support reproducible results. The experiments were performed in a 3.2×9 (metres, width×depth) rectangular area, which represents the dimensions of a typical indoor navigation environment, such as a living room or a small shop. The floor of the environment was dark gray carpet and the boundaries of the environment were made from white curtain material, ensur-

ing participant safety during collisions. A start line was placed parallel to the short side of the room and 0.8 metres from boundary. A finish line was placed 0.8 metres from the opposite boundary. See Figure 7-1 for an illustration of the experimental environment.

Environment contrast and lighting conditions were an important consideration given the nature of the standard intensity visual representation. Upwards facing lights were attached on the roof in order to provide relatively uniform ambient lighting of the scene. Light measurements were taken at the start and end of each experiment to verify that lighting conditions were uniform throughout the course of all experiments.

7.3.3 Overhanging Obstacles

The obstacles used in the study were white cardboard boxes hung from the ceiling. Overhanging obstacles were selected because they are difficult for existing visual aids, such as a cane or a guide dog, to detect. Furthermore, overhanging obstacles such as tree branches and wall-mounted furniture appear frequently in everyday life.

Two sizes of obstacles were used in the study: a 'small' obstacle with dimensions $0.16 \times 0.16 \times 0.16$ (metres, width×depth×height) and a 'large' obstacle with dimensions $0.26 \times 0.26 \times 0.16$ (metres, width×depth×height). The number of obstacles was set to be between 4 and 6, with a random number of up to 2 small obstacles and the remainder large obstacles. This range was selected to provide a challenging setup.

Obstacles were placed in a grid of 7×4 (width×depth) possible locations within the environment. The grid density was chosen so that the participant could not expect to walk in a straight line from the start to the end of the room and avoid all obstacles. At least one obstacle was placed between the participant start position and the centre of the finish line, requiring the participant to adjust their initial trajectory at least once during traversal. All remaining obstacles were placed at random grid locations in the environment. Obstacles were hung at head height for each participant. Figure 7-2b shows two obstacles in the experimental environment, and Figure 7-1 depicts the set of possible obstacle positions in the environment.

7.3.4 Experiment Procedure

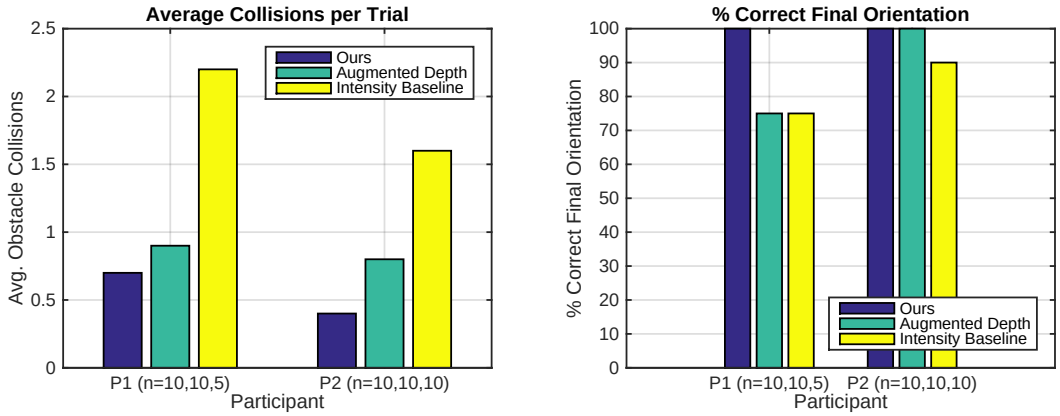
Before each trial, the participant was instructed to wait outside the environment while the obstacles were arranged. At the start of the trial, the participant was led to the centre of the start line and oriented towards the opposite end of the room. The experimenter signalled the participant when the scene representation was turned on and the trial was ready to begin. After this, the participant said the keyword "go" to begin the trial. The experimenter recorded collisions between the participant and the environment during traversal. When the participant crossed the finish line, the experimenter said the keyword "stop" to end the trial.

7.3.5 Evaluation Metrics

Navigation performance during the study was measured through the number of collisions and final participant orientation. These metrics capture task performance in a manner relevant to safe and effective navigation, and are described in detail below.

1. **Number of collisions per trial**: A collision is defined as a contact between the participant, or anything worn by the participant, and the walls and obstacles. The number of collisions primarily reflects the ability of a scene representation to convey obstacle locations to the user. In addition to this, it also captures the ability of the scene representation to facilitate self orientation and route planning, since it is observed that poor self orientation and route planning often results in collisions with the boundary of the environment and even previously bypassed obstacles. This metric is directly relevant to everyday visual prosthesis use, where avoiding collisions is essential for performing safe navigation. The number of collisions per trial is thus the primary metric for this user study.

2. **Final participant orientation**. This records whether the participant was facing the correct boundary of the room at the conclusion of the trial. This metric aims to measure the ability of a visual representation to support self orientation within the environment, differentiating cases where participants correctly nav-



(a) (b)

Figure 7-3. Raw data from the pilot simulation study comparing our surface irregularities visual representation with the state-of-the-art Augmented Depth scene representation and a baseline representation conveying scene brightness, where 'n' denotes the number of trials performed. Two normally-sighted, blindfolded participants performed the task of moving through a corridor while avoiding randomly placed low-contrast overhanging obstacles. The baseline is purely intensity based, to investigate appropriate methods for drawing attention to structurally important features of the environment when they lack visual contrast. Our method results in a lower number of average collisions per trial.

igate to the end of the environment from cases where participants reached the end boundary by chance after losing their orientation within the environment.

Note that the results of the study are reported primarily through the mean statistic of each metric. No further statistical analyses are performed since this study is primarily a pilot study to verify the suitability of surface irregularities for the navigation task, and the raw data and means from the results are sufficiently informative for this purpose. All trial results are processed and plotted using Matlab.

7.4 Results

Surface irregularities resulted in fewer collisions (mean = 0.55) than both Augmented Depth (mean = 0.85) and the standard intensity representation (mean = 1.8), as shown in Table 7.1 and Figure 7-3a. This demonstrates that conveying general structural information provides measurable benefit for orientation and mobility with a prosthetic vision display. In particular, directly conveying scene structure through

	Collisions Per Trial	% Correct Final Orientation
Surface Irregularities	**0.55**	**90**
Augmented Depth	0.85	80
Intensity	1.8	70

Table 7.1. Mean statistics of collisions and final head orientation of all participants from the pilot simulation study, comparing our surface irregularities visual representation with the state-of-the-art Augmented Depth visual representation and a baseline representation conveying scene brightness.

regions of irregular surface shape provides the user with an explicit map of their surroundings, which is more robust and in general easier to understand than scene appearance, and more descriptive than the obstacle locations from Augmented Depth.

Surface irregularities and Augmented Depth both increase contrast on regions that exhibit important structure, making the representations easier to interpret than the intensity representation considering the dynamic range of the display. However, Augmented Depth only increases contrast on ground obstacle locations, and therefore does not directly convey further structural information, such as the corner of a room or other boundaries, to the user. Consequently, participants were more likely to lose their orientation and brush the walls when using Augmented Depth, as evidenced by the higher mean collision rate (mean = 0.85) compared to the surface irregularities representation (mean = 0.55) in Table 7.1.

Navigation using the standard intensity representation resulted in the lowest percentage of correct orientations (mean = 70%), as seen in Table 7.1 and Figure 7-3b. This is due to a key limitation of standard approaches: the reliance on contrast in the environment and increased difficulty in perception when scene contrast is low. In addition to the challenge of discerning low contrast obstacles with this representation, participants experienced difficulty performing self-orientation within the environment due to a lack of high contrast landmarks. Because of this, orientation towards the target boundary was sometimes lost, resulting in traversal into one of the side boundaries of the environment. This problem did not occur when using the surface irregularities representation because participants were able to estimate their orientation from structural cues such as room boundaries.

Trial#	Vis. Rep.	Collisions			Final Orientation
		Wall	Obstacle	Total	
1	Aug	1	0	1	1
2	Aug	4	2	6	0
3	Aug	2	1	3	0
4	Aug	0	1	1	0
5	Aug	0	0	0	1
6	Int	0	5	5	0
7	Int	0	2	2	1
8	Int	0	3	3	1
9	Int	0	0	0	1
10	Int	0	1	1	0
11	Sal	0	0	0	1
12	Sal	2	1	3	1
13	Sal	2	1	3	0
14	Sal	2	0	2	1
15	Sal	1	0	1	1
16	Aug	0	1	1	0
17	Aug	0	2	2	1
18	Aug	0	2	2	1
19	Aug	0	0	0	1
20	Aug	0	0	0	1
21	Sal	1	1	2	1
22	Sal	0	1	1	1
23	Sal	0	3	3	0
24	Sal	1	0	1	1
25	Sal	0	0	0	1

(a) Participant 1

Trial#	Vis. Rep.	Collisions			Final Orientation
		Wall	Obstacle	Total	
1	Aug	0	1	1	1
2	Aug	0	2	2	1
3	Aug	2	1	3	1
4	Aug	0	0	0	1
5	Aug	0	1	1	1
6	Sal	0	1	1	1
7	Sal	0	0	0	1
8	Sal	0	1	1	1
9	Sal	0	0	0	1
10	Sal	0	2	2	1
11	Int	0	2	2	1
12	Int	0	0	0	1
13	Int	0	0	0	0
14	Int	0	1	1	1
15	Int	0	0	0	1
16	Int	0	2	2	1
17	Int	0	4	4	1
18	Int	0	3	3	1
19	Int	0	1	1	1
20	Int	0	3	3	1
21	Aug	0	0	0	1
22	Aug	0	1	1	1
23	Aug	0	0	0	1
24	Aug	0	2	2	1
25	Aug	0	0	0	1
26	Sal	0	0	0	1
27	Sal	0	0	0	1
28	Sal	0	1	1	1
29	Sal	0	0	0	1
30	Sal	0	0	0	1

(b) Participant 2

Table 7.2. Raw collision and orientation results for the pilot study.

7.5 Discussion

7.5.1 Surface Irregularities Clinical Trial

Based on the results of the pilot study, surface irregularities was tested in a clinical trial run by Barnes *et al.* [13] using the same environment and with a similar experimental procedure as the pilot study, apart from five modifications:

- The exclusion of the Augmented Depth experimental condition, since surface irregularities indicated improved performance in the pilot study, and sufficient testing of both conditions for overhanging obstacle avoidance was not possible due to time constraints.

- Inclusion of the system-off baseline, which is a control representation where no visual stimuli is conveyed through the retinal implant, with brightness function given by: $B_{\text{off}}(\varphi) = 0$.

- Inclusion of the preferred visual aid baseline, which measures user performance with their preferred visual aid, in this case a guide dog for both participants.

- Scene representations were presented in blocks of four trials with the same visual representation, in order to ease the cognitive overhead of frequently switching representations.

- The number of trials was increased, with the aim of achieving statistical significance from the study results.

Two retinal implant users with the prototype BVA 24-channel suprachoroidal array took part in the clinical trial, who will be referred to as P1 and P2. The results from these clinical trials provide further insight into the effectiveness of the surface irregularities representation, and will be discussed below.

For both participants, the surface irregularities scene representation resulted in significantly less collisions than the standard intensity representation. This verifies the findings of our simulation study, demonstrating that regions of irregular surface

structure provide important information for navigation decisions during practical use of a retinal implant display.

Surface irregularities was found to result in significantly fewer collisions than system-off for P1, with results for P2 trending towards the same outcome. However, due to time and health considerations, the number of trials involving P2 was fewer than for P1, and insufficient to obtain statistical power on this result.

The study also found that navigation with surface irregularities for P1 resulted in significantly fewer collisions than when using a guide dog, while there was no significant difference in the number of collisions for P2. This is the first recorded instance where use of a prosthetic vision device has led to better performance than with the user's preferred visual aid on a navigation task. The intensity visual representation resulted in worse performance than when using the preferred visual aid for both participants.

7.6 Chapter Summary

In this chapter we have presented an evaluation of our surface irregularities visual representation for navigation with prosthetic vision. Our method reduced the average number of collisions in a pilot navigation study compared to the standard intensity representation and augmented depth for normally-sighted participants using SPV. These results are consistent with results from a subsequent clinical trial involving two retinal implant users, in which use of surface irregularities was found to result in significantly fewer collisions compared to the standard intensity representation. In this clinical trial, surface irregularities also led to equal or better performance than the participant's preferred visual aid. Unlike previous methods, our visual representation provides a depiction of general structural saliency that is robust to appearance. This demonstrates that structural saliency plays an important role in informing navigation decisions during practical use of a prosthetic vision display.

Chapter 8

User Study Evaluation: Bimodal Visual Representation

The previous chapter has shown that conveying salient structure facilitates improved navigation performance with a prosthetic vision display, validating the approach taken in this thesis of employing structural saliency for scene representation. This indicates that conveying fundamentally different types of information can offer distinct advantages when performing a task such as navigation. In particular, we have demonstrated that structure-based cues tend to be more robust and descriptive of the scene when performing navigation, whereas existing work has found that intensity-based scene representations are useful for certain tasks such as viewing navigation symbols or orienting towards high-contrast landmarks in an environment [65, 62]. Therefore, the flexibility of a scene representation could be improved by combining multiple methods into a single display, making available the advantages offered by each individual method. In this chapter we begin to explore this idea by investigating whether structural information can be used effectively when combined with intensity information in a prosthetic vision display. This investigation is performed through a user study that aims to measure the effectiveness of a simple bimodal display compared to standard baseline methods.

Specifically, this chapter presents a novel bimodal visual representation for prosthetic vision called *intensity with cueing*, which combines both structural and intensity

information by mapping a subset of the display to each mode. The proposed visual representation is evaluated through a SPV user study on an orientation and mobility task, which primarily aims to determine whether users are able to employ simultaneously presented structural and intensity information for effective task completion. An introduction is given in Section 8.1. The bi-modal representation is described in Section 8.2, and the task and environment are presented in Section 8.3. The results are given in Section 8.4 and discussed in Section 8.5. The chapter is concluded in Section 8.6.

8.1 Introduction

Structure-based scene representation is necessary for ensuring safe navigation in environments where contrast is low [104]. However, intensity information about the environment is also helpful for performing various tasks related to navigation [65]. For example, intensity information would be required to interpret navigation related symbols such as an arrow painted on a wall. It can also support self-orientation through the process of locating a known high-contrast landmark, such as a door, and using it as a reference point during navigation. Therefore, combining structure-based cues with scene intensity information can offer improved flexibility during prosthetic vision navigation scenarios.

In this chapter, we seek to determine whether a bimodal scene representation can be effectively used to perform navigation. We first propose a novel bimodal vision processing method called intensity with cueing, which conveys both intensity and structural information to the user. The design of such a representation is challenging compared to the design of existing scene representations, due to the extra information from two different modes of information that must be summarised within the already limited display capacity of the device. We investigate the feasibility of a bimodal representation through a simple design. We take the approach of partitioning the display into a set of phosphenes that convey intensity information, and a set of phosphenes dedicated to conveying structural obstacle cues. Specifically, our approach modifies

a standard downsampled intensity display so that the bottom row of phosphenes conveys the presence of obstacles on the ground, as shown in Figure 8-1.

Previous work has investigated the inclusion of saliency cueing on a standard intensity display [112]. However, this method uses appearance-based saliency, and can not robustly detect the presence of low-contrast obstacles during navigation. Furthermore, the directional cues are designed to be on-demand, and requires the user to explicitly initiate the saliency detection process each time saliency guidance is desired. Therefore, this method is not ideal for performing safe navigation, where obstacle detection should occur automatically and continuously. Unlike previous work [112], intensity with cueing simultaneously provides real time display of both scene intensity and robust structure-based cues, facilitating safe navigation through improved perception of obstacle structure and scene appearance.

In order to determine the feasibility of a bi-modal representation, we perform evaluations through a user study in which participants use SPV to perform an orientation and mobility task. The task involves navigation towards a high-contrast landmark through a room-like environment containing low-contrast obstacles placed on the ground. This study aims to test whether users are able to perform a complex navigation task using the two different modes of information, and whether the availability of condensed structural information on an intensity-based display improves performance. Thus, the experiment is run as a comparative study, with the performance of intensity with cueing being compared against the standard intensity representation and a control condition. In order to more effectively estimate implant user performance, we base our SPV display on an irregular spatial layout constructed from participant-reported phosphene locations. These phosphene locations were recorded as part of initial clinical trials of a 24-channel prototype implant [7].

This chapter thus provides the following contributions: evaluation of a novel bi-modal visual representation for navigation with SPV; demonstration of competency in an orientation and obstacle avoidance task with less than 20 phosphenes with SPV; and, simulation using a layout from implanted participant reported phosphene locations. Our results provide new insights into how vision processing could be used

to alleviate display constraints with current and near-term retinal prostheses and improve orientation and mobility outcomes for retinal prosthesis users.

8.2 Intensity with Cueing Representation

This section introduces intensity with cueing, a bimodal visual representation for orientation and mobility that combines image intensity and depth-based obstacle cues. Like Augmented Depth [103], this representation explicitly highlights the presence of ground obstacles in a manner independent of their colour or lighting conditions. However, scene intensity is simultaneously presented in order to facilitate self-orientation using intensity-based cues such as light sources and high-contrast landmarks. These two modes of information are combined by partitioning the output electrodes into an obstacle cueing set Φ_{obst} and an intensity set Φ_{int}, with stimulation levels computed independently for each set. Figure 8-2 gives an overview of the main computation stages for these two output sets.

The obstacle cueing set $\Phi_{\text{obst}} = \{\phi_1, \phi_2, ..., \phi_n\}$ comprises the bottom row of electrodes in the implant. Each electrode aims to convey the presence of obstacles within a fixed input field R_i of the visual field. For the purposes of ground obstacle avoidance, these regions are defined by vertically splitting the bottom half of the camera visual field into equally sized rectangles $R_1, R_2, ..., R_n$. This ensures only the most relevant information for ground obstacle avoidance is conveyed when the user is forward looking, as shown in Figure 8-1. The order of these input fields is consistent with the horizontal ordering of the phosphenes in Φ_{obst}, *i.e.* , the left-most phosphene corresponds to the left-most region and so on. Thus, the input fields form a polar histogram centered on the depth sensor, which is visualized by the obstacle cueing phosphene set. Polar obstacle density histograms are also used in the vector field histogram (VFH) algorithm [18], a well established method for robot obstacle avoidance that has previously been incorporated into audio and tactile mobility aids for the vision impaired [134]. However, we additionally modulate the display of obstacle distance by relative image height, with a higher height implying that the obstacle is

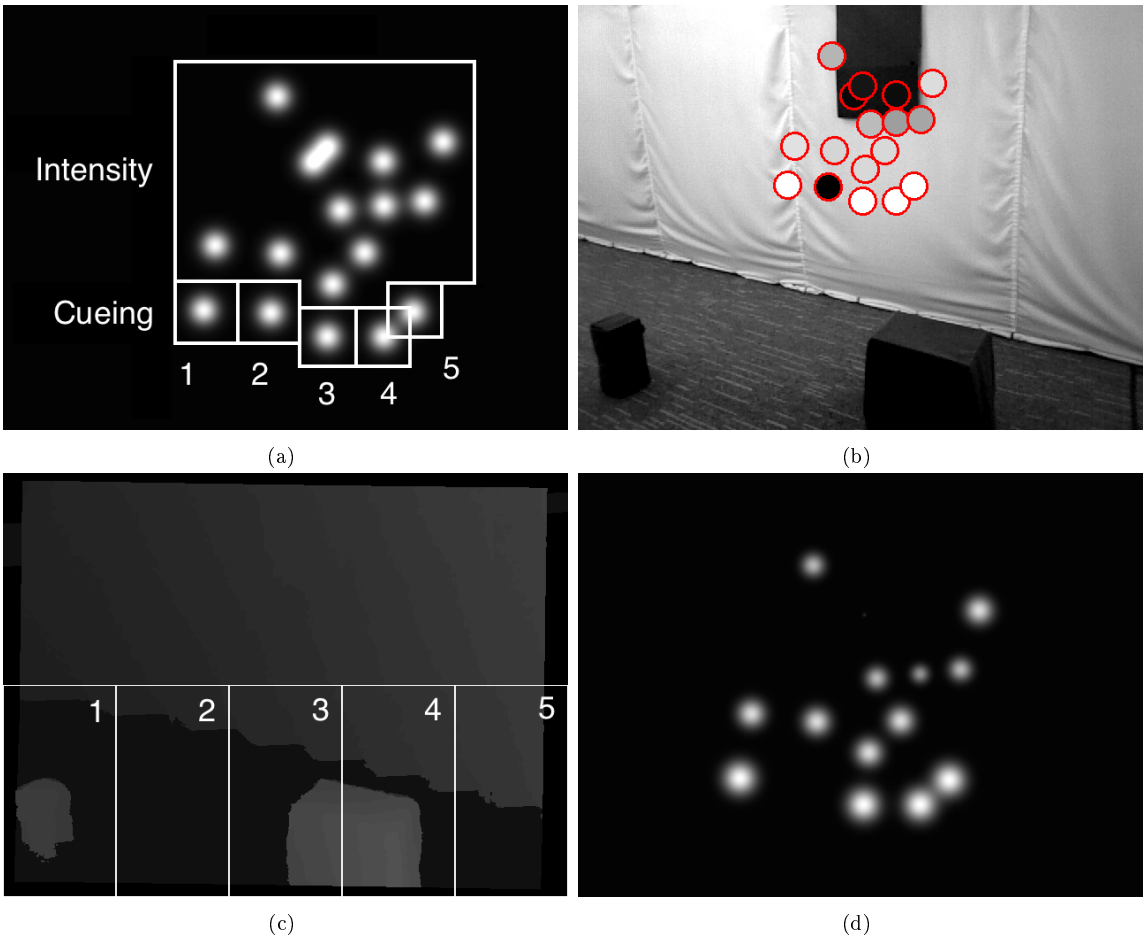


Figure 8-1. The intensity with cueing visual representation. (a) Simulated prosthetic vision of user-reported phosphene layout, showing the partition of the phosphenes into the intensity set Φ_{int} and the obstacle cueing set Φ_{obst}. (b) Intensity image with projected phosphene locations circled in red and showing the brightness level of each phosphene. (c) Obstacle disparity image with detection regions, each region is labelled with the corresponding output cueing phosphene. (d) Simulated prosthetic vision of the scene, in which the black target is visible as the dark region near the center top of the display. Dark obstacle cueing phosphenes, such as ϕ_2, indicate that the corresponding direction of travel is free of obstructions. Note that when the user's head is level, the intensity set is positioned for viewing landmarks at eye height while the obstacle cueing set indicates nearby obstacle presence on the ground.

further away, in order to amplify differences in obstacle distance in a low dynamic range display.

Obstacle presence is conveyed by varying the stimulation level of each cueing phosphene ϕ_i according to the disparity values of pixels in R_i, with larger disparity values implying closer surfaces and therefore producing a higher response. Thus, the input depth map D is first converted to a disparity map $\mathcal{D}$. Use of raw disparity values in this process is not ideal, since there is often little depth difference between trip hazards and the surrounding ground. In order to obtain a more useful obstacle response, we first apply a segmentation algorithm to identify the set of pixels $\mathcal{O}$ that correspond with obstacle surfaces in the disparity image, and then map all other pixels to zero, giving the obstacle disparity image D_{obst}:

$$D_{\text{obst}}(p) = \begin{cases} \mathcal{D}(x,y), & \text{if } (x,y) \in \mathcal{O} \\ 0, & \text{otherwise.} \end{cases} \tag{8.1}$$

Thus, all non zero disparities in D_{obst} correspond to an obstacle surface. In our implementation, we use a ground-plane segmentation algorithm [101] to estimate $\mathcal{O}$, which results in all non-ground pixels being considered obstacles.

The stimulation level of each cueing electrode is obtained by accumulating the disparities within the corresponding input field of D_{obst}. However, abrupt boundaries of the input fields can lead to aliasing in the output display, reducing the accuracy of obstacle localization. Therefore, we incorporate an anti-aliasing function $J(x - R_{i,x})$ to modulate the disparity of a point (x, y) based on the horizontal distance $x - R_{i,x}$ from the center of the region.

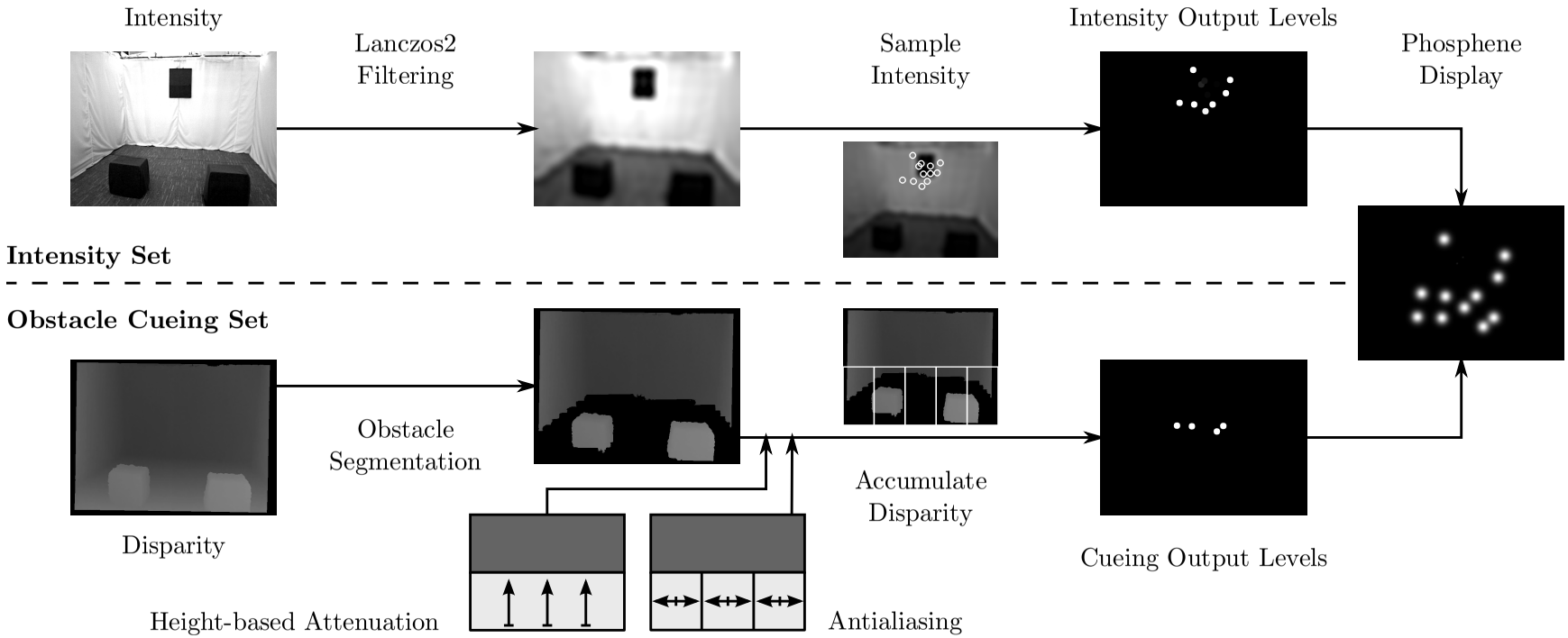


Figure 8-2. Overview of the main computation stages of our method, which separately processes the intensity and obstacle cueing electrode sets and combines them in the final display. The intensity set output levels are computed by sampling a Lanczos2 filtered intensity image at the projected phosphene locations. The obstacle cueing set output levels are computed by accumulating segmented obstacle disparity values that are attenuated by pixel height and anti-aliased for visualization. Phosphene placement will be described in Section 8.3.2.

Vertically, aliasing is not an issue since there are no boundaries between regions to consider. However, unlike Augmented Depth where vertical position also indicates object depth, the proposed approach encodes obstacle distance using only a single value. Given the poor resolution of relative distance between nearby obstacles in different R_i due to the restricted dynamic range of phosphenes, we allow the user to amplify this signal through their head tilt. This is achieved by attenuating non-ground plane disparity by its image height y with a function $K(y)$. Figure 8-3 shows examples demonstrating the effect of antialiasing and height-based attenuation.

Thus, accumulating the anti-aliased and attenuated disparities of the input field gives the output level:

$$B_{\text{cue}}\left(\phi_i, D\right) = \int_{(x,y)\in R_i} D_{\text{obst}}(x,y) \cdot J\left(x - R_{i,x}\right) \cdot K(y) \tag{8.2}$$

In the implementation, as the size of the R_i are large, we use a linear approximation for filtering. Specifically, J and K are implemented as piecewise linear functions that peak at the center of the input field and the bottom of the image respectively.

We assign the standard intensity brightness function B_{lanczos2} defined in Equation 2.14 to phosphenes in the intensity set. Thus, the brightness function for our intensity with cueing visual representation is given by:

$$B_{\text{intcued}}(\phi, I, D) = \begin{cases} B_{\text{cue}}(\phi, D), & \text{if } \phi \in \Phi_{\text{obst}} \\ B_{\text{lanczos2}}(\phi, I), & \text{otherwise} \end{cases} \tag{8.3}$$

8.3 Experiment Design

The proposed visual representation was evaluated in a SPV study on an orientation and mobility task. Eight volunteers (five female, three male, aged between 20 and 34 years) with normal or corrected-to-normal vision of at least 20/20 took part in the study. Informed written consent was obtained from all participants before they began the experiment. The ethical component of this research was approved by the

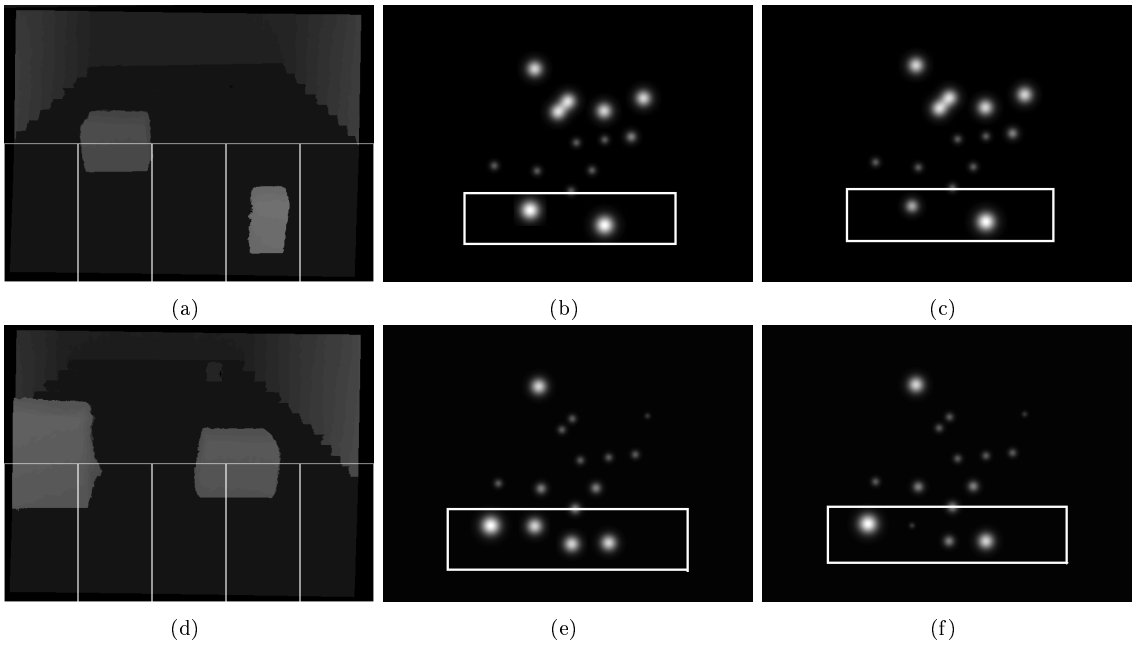


Figure 8-3. Top row: illustration of horizontal antialiasing, which improves obstacle localisation. (a) obstacle disparity image with detection regions marked; (b) phosphene rendering without horizontal antialiasing; (c) with horizontal antialiasing. Bottom row: illustration of vertical disparity attenuation, which amplifies the relative depth between obstacles. (d) obstacle disparity image with detection regions marked; (e) phosphene rendering without vertical attenuation; (f) with vertical attenuation. The obstacle cueing sets are marked with a rectangle in the phosphene images.

Australian National University Human Research and Ethics Committee. The study adhered to the tenets of the Declaration of Helsinki.

Navigation performance was measured on the task of moving towards a high-contrast target while avoiding low-contrast obstacles placed in the environment. This task is an adaptation of the walk-to-door task used in the clinical trials of the Second Sight Medical Products Argus II implant [65]. Notably, we introduce obstacles into the traversable space, reduce the size of the target, and increase the size of the experimental environment.

This experiment employed a randomized control design that consisted of repeated measures of all visual representations for each participant. Each participant performed 15 trials with the intensity representation, 15 using intensity with cueing, and 5 with random, presented in a random order.

8.3.1 Visual Representations

The performance of three vision processing methods was measured in this study: intensity with cueing, the standard intensity method, and a randomised control condition. Details of each visual representation are presented below.

- Intensity with cueing: the proposed bimodal scene representation proposed in this chapter, which incorporates structural obstacle cues into a standard intensity display.

- Intensity: the standard downsampled intensity representation. To reflect current state-of-the-art, downsampling is performed using Nyquist band-limited Lanczos2 filtering. The brightness function for this representation is given in Equation 2.14. Figure 4 shows an example of a scene rendered using the intensity representation.

- Random: A control condition in which the stimulation level of each electrode is set randomly every frame. The brightness function of this representation is given by $B_{\text{rand}}(\phi) = r$, where r is a random value within the dynamic range of ϕ. Random gives no useful information to the user, but unlike a blank screen, is not immediately identifiable to participants, thereby controlling against any potential bias. The use of randomized representations to measure 'device off' performance is becoming standard practice in the clinical evaluation of visual prostheses [104].

8.3.2 Mobile SPV System

Participants used a wearable mobile simulation system to provide a real time prosthetic vision simulation of the environment during the study. This system performed vision processing on the output of a head mounted RGB-D camera and rendered the resulting phosphene simulation to a head mounted display. For more details on the system hardware, see Chapter 7.

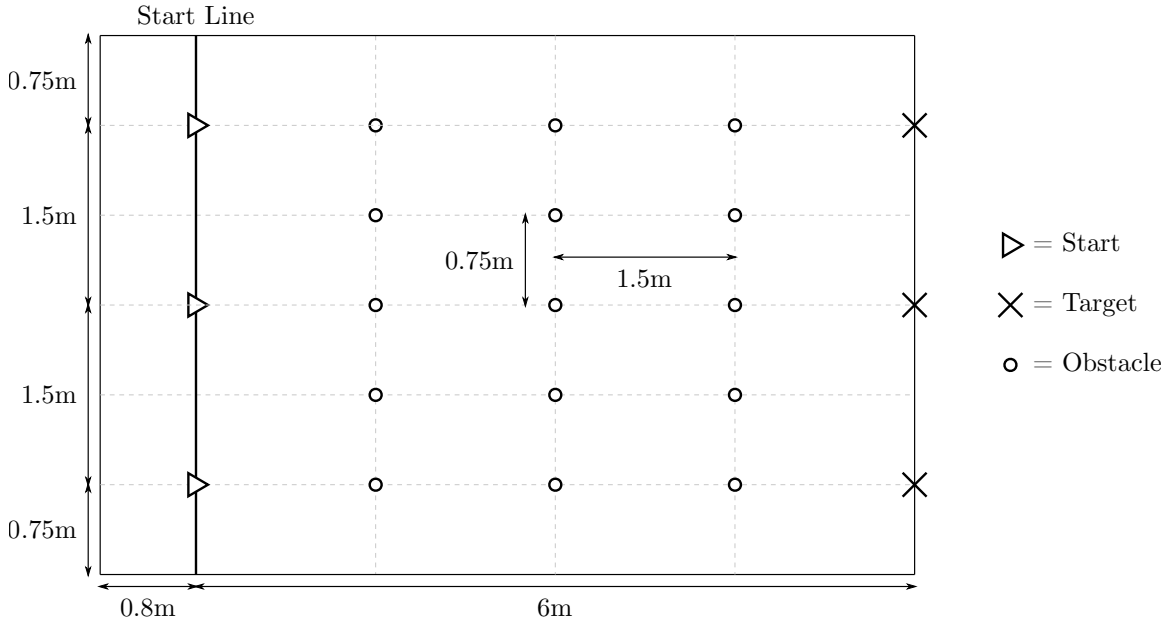


Figure 8-4. A top-down diagram of the experimental environment. The possible participant start positions are marked as triangles, possible obstacle locations are denoted with circles, and potential target locations are denoted with a cross.

Phosphene positions in the simulated display were based on the recorded location of individual phosphene locations as reported by a patient with a prototype 24-channel suprachoroidal implant. Thus, the displayed image contained 17 non-overlapping phosphenes rendered at irregular locations within a 15×15 degree field of view. Phosphenes were rendered as circular Gaussian spots with 8 output levels, using the implementation described in [104].

8.3.3 Navigation Environment

The experiment took place in a purpose-built 6.8×4.5 (metres, width×depth) rectangular space with dark grey carpeted floor and walls constructed from white curtain material. Upwards-oriented lamps were installed at regular intervals in the ceiling to provide uniform lighting conditions across trials. The trial start line was located 0.8 metres from and parallel to the width of the environment. The participant start location for each trial was randomly selected from three equally spaced start positions along this start line. See Figure 8-4 for a diagram of the navigation environment.

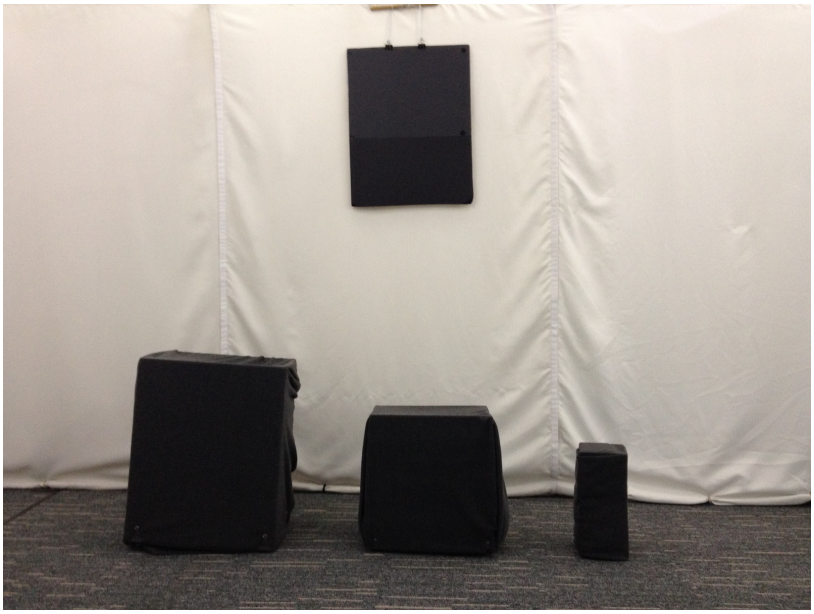


Figure 8-5. A view of the navigation target and the three different obstacle sizes used in the study.

8.3.4 Navigation Target

During the trial, participants aimed to move towards a navigation target placed in the environment. The navigation target was a 0.54×0.7 (metres, width×height) cardboard rectangle, coloured black in order to have high-contrast with the surrounding white wall. The target was placed at one of three equally spaced target locations on the opposite wall to the start line, at a fixed height of 1.7 metres from the ground. Figure 8-5 shows the navigation target used in the study.

8.3.5 Ground Obstacles

The obstacles used in the study were cardboard boxes covered in a soft felt material and placed on the ground. Obstacles were dark grey in colour and thus had low contrast with the ground. Obstacles ranged in size from small ($0.18 \times 0.18 \times 0.38m$ metres, width×height×depth), medium ($0.45 \times 0.3 \times 0.5$ metres, width×height×depth), to large ($0.54 \times 0.34 \times 0.66$ metres, width×height×depth). Figure 8-5 shows the obstacles used in the study.

A random quantity of between 1 and 4 obstacles were placed in the environment during each trial. Each trial had either a low obstacle density (1-2 obstacles present during navigation) or a high obstacle density (3-4 obstacles present during navigation). Obstacle sizes were chosen randomly. The obstacles were placed within a 3×5 (width$\times$ depth) grid of possible locations, with the location of each obstacle selected randomly from this grid (see Figure 8-4).

8.3.6 Training

Familiarization exercises were carried out within the test environment prior to commencement, during which time participants were shown the target using the intensity representation from various distances and angles. Training also focused on familiarization with the different sized obstacles with both the intensity and intensity with cueing representations. Practice trials were performed with both the intensity and intensity with cueing representations until participants were familiarised with the task.

8.3.7 Experiment Procedure

Before each trial, the participant waited outside the environment while the obstacles were arranged. At the start of the trial, the participant was led to the start position and oriented towards the opposite end of the room. Note that the participant was led along a circuitous route to the start position, in order to obfuscate which of the three start positions was being used. The experimenter signalled the participant when the scene representation was turned on and the trial was ready to begin. After this, the participant said the keyword "go" to begin the trial, at which point time recording began. The experimenter recorded collisions between the participant and the environment during traversal. When the participant judged that they had reached the target, they said the keyword "stop" to end the trial, and the time measurement was halted. Participants were instructed to get as close as they could to the target without touching it. The distance between the participant and the target at the

conclusion of the trial was recorded.

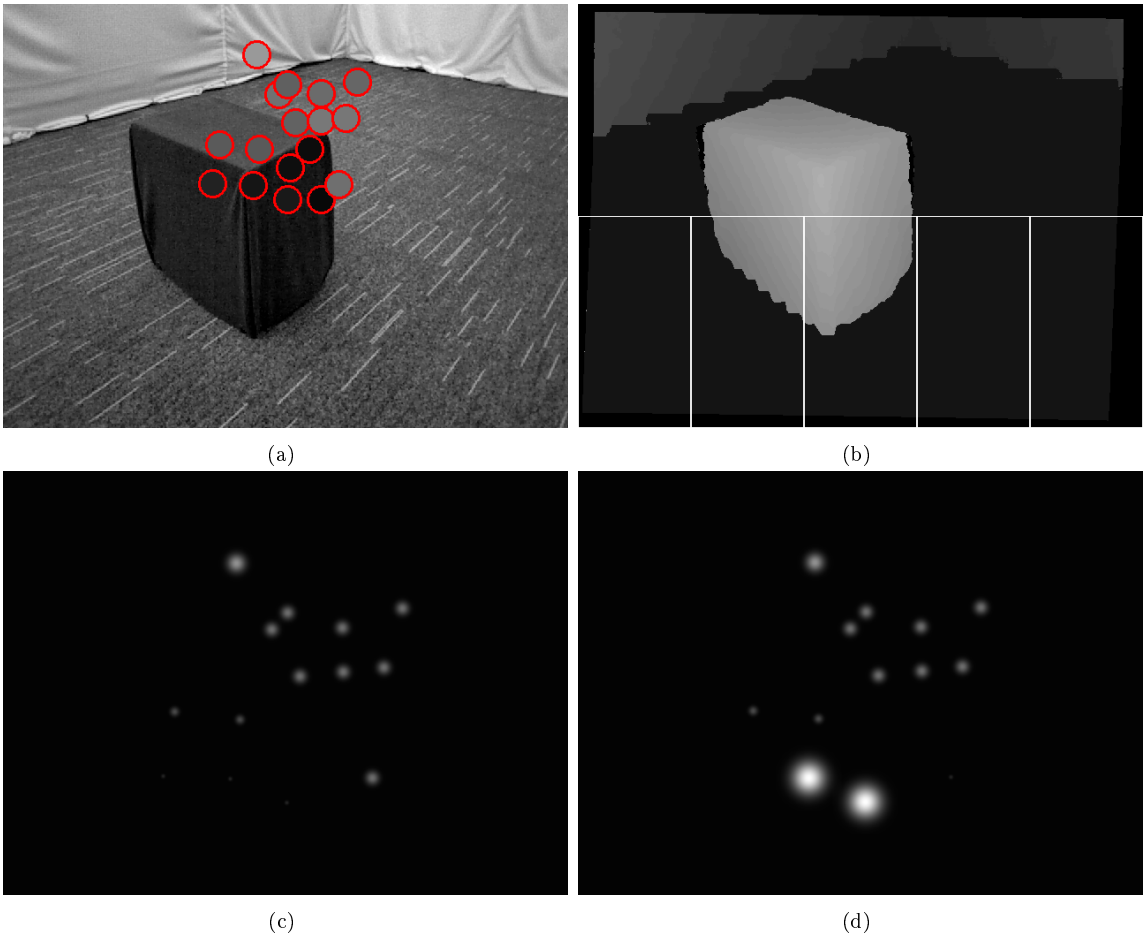
8.3.8 Performance Measures

Participant performance in the navigation task was measured according to three performance measures: number of collisions, final distance from the target, final head orientation, and percentage of preferred walking speed. Each metric is described in detail below.

- **Number of collisions per trial**. A collision was defined as a contact between the participant, or anything worn by the participant, and the experimental environment. This measures the ability of the visual representation to facilitate safe navigation through obstacle avoidance.

- **Final distance from the target** was defined as the distance between the participant and the target at trial completion. Specifically, this was measured as the distance between the center of the line between the front of the participant's shoes, and the ground beneath the center of the navigation target, recorded to a precision of 0.005 metres. This metric reflects the ability of the visual representation to facilitate self orientation based on a high-contrast landmark. Participants that are unable to reach the landmark at the trial conclusion will have a high value for this metric. This also measures the ability of the visual representation to enable visually guided judgement of object distances, since participants aimed to get close to the target without touching it.

- **Percentage of preferred walking speed (PPWS)** is a standard metric used in prosthetic vision studies and low vision rehabilitation training, which measures the walking efficiency of participants when using a visual representation compared to their normal walking speed. Specifically, percentage of preferred walking speed is calculated by dividing a participant's average walking speed per trial by their preferred walking speed. Preferred walking speed for each participant was measured when walking using normal vision while wearing the SPV system.



(a) (b)

(c) (d)

Figure 8-6. Example of rendering scene intensity with and without obstacle cueing: (a) camera intensity image with projected phosphene locations; (b) obstacle disparity image with detection regions; (c) rendered using a standard intensity representation without obstacle cueing; (d) rendered using the intensity with cueing representation.

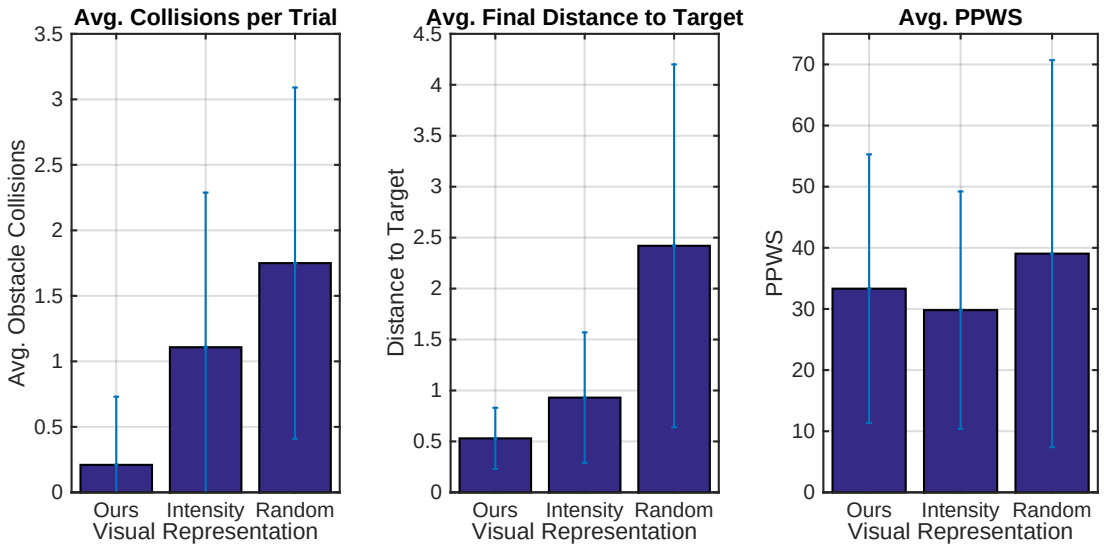


Figure 8-7. Means and standard deviations of performance outcomes for our method (intensity with cueing), the standard intensity representation, and the random visual representation.

8.3.9 Statistical Analysis

We perform statistical analysis to determine the statistical significance of the results of the study. Exploratory data analyses indicated mild positive skewness for the number of collisions which was transformed using square root, while the moderate skewness evident for PPWS and Final Distance from the Target was corrected with a logarithmic transformation. Models of analyses of variance controlled for potentially confounding effects of factors relevant to each performance measure as follows: number of collisions (obstacle density); final distance from the target (participant, trial presentation order); and, PPWS (participant). An effect was considered significant at $p = 0.05$. Comparisons between participants and visual representations for percentage of accurate responses were calculated using χ_2 statistics. All statistical analyses were performed using version 20 of SPSS for Windows (SPSS, Chicago, IL).

Metric	Visual Representation	Number of Trials	Mean	Standard Deviation
Collisions	Intensity with cueing	120	0.21	0.52
	Intensity	120	1.11	1.18
	Random	40	1.75	1.34
	Total	280	0.81	1.13
Distance	Intensity with cueing	117	0.53	0.30
	Intensity	118	0.93	0.64
	Random	40	2.42	1.78
	Total	275	0.98	1.03
PPWS	Intensity with cueing	120	33.32	21.96
	Intensity	120	29.81	19.41
	Random	40	39.05	51.65
	Total	280	32.63	27.36

Table 8.1. Counts, means and standard deviations for orientation and mobility task outcomes in SPV (N = 8).

8.4 Results

8.4.1 Participants

Analysis of variance indicated no significant interaction between participants and the number of collisions per trial with $p = .50$. Based on this, the data from the eight participants were pooled for further analyses to enhance the statistical power. There were, however, significant ($p < .001$) individual differences for Final Distance from the Target and PPWS, which was accounted for in further analyses.

8.4.2 Number of Collisions

Examination of the means (Table 8.1) indicated that the vision processing method significantly impacted on the number of collisions with $F_2 = 52.75, p < .0001$. Direct comparisons were made between intensity with cueing, intensity, and random vision processing methods. Intensity with cueing (mean $= 0.21 \pm 0.52$) was associated with significantly fewer collisions than intensity (mean $= 1.108 \pm 1.18, p < .0001$; large effect size $d = .99$). Both intensity with cueing ($p < .0001$; large effect size $d = 1.52$) and intensity ($p = .001$; moderate effect size $d = .51$) were associated with

significantly fewer collisions in comparison to random (mean = 1.75 ± 1.34) visual representation. No main effect was apparent for obstacle density ($p = .09$) on the number of collisions; however, a significant interaction was found between visual representation and obstacle density with $F_2 = 3.62, p = .028$. Inspection of the means indicated that the intensity with cueing ($p = .30$) and random ($p = .72$) vision processing methods were not significantly impacted by an increase in obstacle density, while an increase in obstacle density lead to a greater number of collisions for the Intensity ($F_1 = 10.91, p = .001$) visual representation. These results are summarized in Figure 8-7a.

8.4.3 Final Distance from the Target

The type of visual representation (*i.e.* , intensity with cueing, intensity, random) had a significant impact on the final distance to the target with $F_2 = 71.32, p < .0001$. Intensity with cueing (mean = 0.53 ± 0.30) was associated with significantly fewer collisions than intensity (mean = $0.93 \pm 0.64, p < .0001$; large effect size $d = 0.80$). Both intensity with cueing ($p < .0001$; large effect size $d = 1.48$) and intensity ($p < .0001$; large effect size $d = 1.11$) were associated with significantly fewer collisions in comparison to the random (mean = 2.42 ± 1.78) visual representation. These results are summarized in Figure 8-7b.

8.4.4 Percentage of Preferred Walking Speed

Examination of the means (see Table 8.1, Figure 8-7c) indicated that there was no main effect for the visual representation (*i.e.* , intensity with cueing, intensity, random) on PPWS with $F_2 = 2.33, p = .099$. Intensity with cueing ($33.32 \pm 21.96\%$ of Preferred Walking Speed) achieved a similar PPWS to the intensity ($29.81 \pm 19.41\%, p = 0.12$) and random ($39.05 \pm 31.65\%, p = 0.054$) visual representation methods. PPWS did not significantly differ for the intensity and random ($p = 0.41$) visual representation methods. The obstacle density did not significantly impact on the PPWS with $p = 0.063$.

8.5 Discussion

This study shows that simultaneously presenting intensity information and obstacle cues in a single visual representation can provide benefit for individuals performing an orientation and mobility task using simulated prosthetic vision. This is the first study evaluating a visual representation that combines two different types of information continuously in a single display, demonstrating that users are able to process and successfully make use of the bimodal display when completing an orientation and mobility task. Whereas previous work predominantly maps stimulation levels exclusively to intensity or depth information, our results indicate that incorporating cues derived from scene understanding techniques along with intensity cues have the potential to improve functional outcomes with visual prostheses.

The mean number of collisions per trial for intensity with cueing was significantly lower than that of the standard intensity representation. It appears evident that the incorporation of cueing provided a clear indication of the presence of obstacles in the test environment. The qualitative comparison of intensity with and without cueing in Figure 8-1 supports this, showing the advantages of structural obstacle detection methods in low-contrast environments. Furthermore, the inclusion of obstacle cueing did not increase the difficulty of display interpretation, evidence by the lack of significant impact on PPWS.

Results obtained using intensity with cueing showed no significant difference in collision rate between trials conducted with high or low obstacle density, which suggests the visual representation is scalable with respect to obstacle density. This follows from the fact that the structure component of intensity with cueing clearly conveys obstacle locations. Conversely, there was a significant increase in collision rate when the Intensity representation was used on high density obstacle courses. This suggests that standard intensity-based representations under the simulated display conditions are less suitable for navigation in the presence of a large number of low-contrast obstacles, which present an intrinsically challenging scenario for these representations given the limited dynamic range of prosthetic vision devices. Indeed, it was observed

during trials that participants often appeared to have difficulty determining obstacle locations when using the Intensity representation, particularly when multiple obstacles were present within the field of view.

Participants achieved closer final distances to the target using intensity with cueing than the intensity representation. The depth-based cueing of oncoming surfaces may have facilitated better judgement of proximity to the surface. Specifically, the direct encoding of surface proximity through the depth-based modulation of obstacle cueing electrode levels is likely to have assisted participants in judging their final desired proximity. Conversely, target distance judgement using the intensity representation was more difficult to infer from the intensity image, with participants needing to estimate the distance by locating the boundaries of the target or the wall-floor boundary, resulting in a likely increase in uncertainty. This highlights a scenario in which scene structure information can be more descriptive than appearance.

Unlike previous studies, the locations of phosphenes in our simulation display are determined directly using reported phosphene locations from an implanted participant. It is well established that retinal implant users generally perceive phosphenes in a distorted, irregular layout that is generally not modeled in simulation studies. The results indicate that our method is robust to the spatial irregularities of a real display, and does not rely on regular structure in the phosphene layout to achieve effective task performance.

Obstacle cueing electrode levels were anti-aliased based on the quantity, disparity value, and image location of obstacle pixels in the input field. Study participants had little difficulty interpreting the cueing set, despite only limited training. We speculate that the anti-aliasing of the phosphenes enabled participants to accurately estimate obstacle locations by interpreting the change in phosphene brightness with respect to head movements, as evidenced by the lower collision rates for intensity with cueing.

An alternative to performing obstacle cueing is to produce a high-level travel direction cue [134]. While a travel direction cue would be potentially be easier to convey through devices with limited output bandwidth, our approach of general obstacle cueing has the advantage of providing a lower level representation of the environment

which allows users to more directly interpret the scene. For example, rather than just knowing which way to walk, it may be more useful to know where obstructions are in the environment. This enables users to make their own decisions on travel direction based on their surroundings, and to build a mental map of the environment as they travel. Furthermore, the lower level information from our cue provides more information for making navigation decisions when combined with data other mobility aids.

8.6 Chapter Summary

In this chapter we have presented a bimodal scene visualization technique for orientation and mobility with prosthetic vision, which simultaneously conveys scene intensity and structure-based obstacle cues in a single display. The obstacle cues are computed from a histogrammed image of obstacle disparities, with antialiasing and height-based disparity attenuation applied to enhance obstacle localization. The representation was evaluated in a simulated prosthetic vision study, resulting in improved performance on an orientation and mobility task compared to the standard intensity representation prevalent in the literature. This provides further validation for the application of sophisticated vision processing and scene understanding methods as a means of alleviating display constraints for near term prosthetic vision devices.

Our results show that the presentation of structure-based cues in addition to intensity information can support navigation with a prosthetic vision display. Users are able to interpret the simultaneous presentation of both types of information, and exploit the advantages of each to more effectively perform the task.

This work validates the approach of merging different types of cues onto a single display. With the increasing interest in and development of novel vision processing methods, our approach provides a solution to combining the advantages offered by different vision methods for a target application in a single display.

Chapter 9

Conclusion

9.1 Summary of Thesis Findings

This thesis has demonstrated how knowledge of scene structure can inform visual representations for prosthetic vision. The approach of this thesis has established a novel paradigm for scene representation based on modelling general salient structure, unlike previous approaches which predominantly focus on intensity or limited structural obstacle cues. This thesis has also introduced a set of novel methods for detecting two key types of salient structure. Specifically, we have investigated salient edge structure for conveying the general shape of a scene, and salient object structure for modelling visual attention based on normal human visual perception. We now summarize the thesis findings for addressing these two problems, as well as the thesis findings from evaluation on the prosthetic vision application.

9.1.1 Salient Edge Detection

We have made the observation that regions of structural change are informative for conveying the shape of a scene. Surface regions with locally irregular shape, such as room boundaries, clutter, and trip hazards, convey the general shape of a scene and are important for performing physical tasks such as navigation. On the other hand, featureless surfaces with uniform shape such as walls and the ground are relatively un-

informative. Therefore, this thesis has proposed a structural saliency measure based on surface irregularity to convey general scene shape for prosthetic vision navigation. Surface irregularity was measured through multi-scale analysis of iso-disparity contours, and was found to more reliably detect relevant surface boundary structure than existing structural scene representations for prosthetic vision.

We have noted that while surface irregularities was able to identify the salient structure in an image, the precision of the method was limited by the low level nature of the detection. For example, ripples in a curtain should have a lower saliency than a perpendicular join between two walls, however these two cases have a similar amount of change in local surface shape and therefore can be difficult to distinguish using low level analysis. We thus extended our approach by learning high level information based on human judgements of structural importance in order to better distinguish salient edge regions. A deep learning framework was used to extract this high level information from annotated depth images. In order to facilitate the learning process, we have introduced the novel DSD depth input encoding. The DSD encoding was designed to provide a minimal encoding of the depth image that captures the fundamental geometric properties of a surface. Our salient edge detection system was evaluated against state-of-the-art systems on a new salient edge dataset, containing a variety of prosthetic vision scenarios in which boundaries important for understanding the scene have been labelled. Compared to the standard methods, DSD was able to better characterise the different types of structure that correspond to salient edges. Thus, our method enables improved detection of structure that is important for understanding a scene.

9.1.2 Salient Object Detection

Accurate modelling of regions or objects that are salient to the human visual system enables vision processing methods to more closely emulate the function of biological saliency detection. While most structural salient object detection methods assign saliency based on depth contrast, we have proposed that surface shape is an important factor for determining what makes an object salient. We have introduced a method

to quantify surface shape through the HOSO feature, which captures the distribution of surface normals within an image region. Saliency was thus computed by measuring the contrast between the HOSO features of different surface patches. Our method was found to more accurately predict structurally salient object locations than competing methods, demonstrating that regions with contrasting surface shape are likely to be salient to the human visual system.

We observe that HOSO analysis and contrast-based methods in general have a number of limitations for application to prosthetic vision scenarios. Most notably, these methods tend not to reflect any consistently meaningful physical quantities within the scene, since contrast is largely dependent on random factors such as object placement and viewpoint. In order to address this issue, we have developed a new formulation of saliency that captures a more stable and descriptive measure of structural importance. Our proposed LBE method thus measures saliency based on the degree to which a region is surrounded by local background. Our raw LBE feature was found to detect structurally salient regions with high accuracy compared to existing depth contrast methods. When combined with standard saliency components to form a salient object detection system, our method produced better results than state-of-the-art systems at the time of contribution. These results demonstrate that background enclosure accurately models scene structure that is salient to the human visual system.

9.1.3 Evaluation for Prosthetic Vision

We have proposed a visual representation based on the surface irregularities formulation of salient structure. In order to gauge the effectiveness of our structure-based representation during practical use of a prosthetic vision display, this visual representation has been evaluated through a user study on a navigation task. This was a crucial step in determining whether surface irregularities could be applied to effectively achieve the salient edge detection goal of conveying scene shape. The results of the study demonstrated that our method was associated with improved performance compared to standard methods, implying that users were able to better interpret

scene shape using surface irregularities and thus plan their actions to more effectively achieve task objectives. Furthermore, we have demonstrated that conveying general scene shape through surface irregularities resulted in better performance than conveying only structure-based obstacle cues as in previous methods [103], since the scene structure information could be used to perform actions that support navigation such as self-orientation and building a mental map of the environment. We concluded from these results that surface irregularities effectively conveys scene shape for understanding a scene when performing navigation with prosthetic vision.

We have observed that many different types of vision processing methods have been proposed, each offering advantages for performing certain tasks. We investigate a new method of scene representation that combines the advantages of different types of information by merging them on the display. We test the feasibility of this approach by evaluating a novel bimodal representation that simultaneously conveys both intensity and structural information on a single display. Evaluation is performed through a user study on an orientation and mobility task. The bimodal representation was found to result in improved obstacle avoidance performance with no difference in walking efficiency compared to standard methods. This provides insight on the design of future scene representations, demonstrating that multiple modes of information can be interpreted from a single display to effectively perform a task.

9.2 Limitations and Future Work

This thesis has provided the foundation for structure-based scene representations for prosthetic vision, and initiated many directions for future work. However, the work presented also has a number of limitations.

- We have demonstrated the effectiveness of the background enclosure formulation of saliency through the LBE feature. The implementation of background enclosure based saliency detection could be refined by learning high level semantic, scale, and context information from an LBE base representation. In particular, deep CNN architectures show promise for extracting this type of

information and improving the saliency detection results. Sufficient training of such a system could resolve many boundary cases such as the failure cases shown in Figure 6-12.

- The methods and evaluations in this thesis have been presented in the context of navigation tasks, since navigation is a particularly challenging task which benefits from structural information. Future work can examine the contribution of the proposed structural methods for performing other physical tasks, such as tabletop tasks and grasping tasks, with prosthetic vision.

- While our salient edge detection system has been shown to provide improved performance on a dataset of static images related to prosthetic vision, it has yet to be tested in a live navigation environment. Similarly, the salient object detection methods have only been evaluated on standard datasets, and have not been evaluated for performance of prosthetic vision tasks. This was largely due to the tremendous overhead of performing a user study, which limited the number of studies that could be run. Future work would focus on further evaluation of these methods on prosthetic vision tasks. For example, the deep salient edge method from Chapter 4 could be evaluated on a navigation task, in order to estimate the improvement offered by incorporating high level information to refine the detection process. Additionally, our salient object detection methods could be applied to perform structure-based orientation, object search, or tabletop tasks.

- We incorporate task-specific cues into our bimodal visual representation by partitioning the electrode display into a cueing set comprising the bottom row of electrodes, and an intensity set composed of the remainder of the display. While this approach is appropriate for ground obstacle avoidance, it does not readily extend to other activities of daily living. Subsequent work by McCarthy *et al.* proposes a more generalized method of ensuring task-relevant features are visible in an intensity-based display using augmentations of contrast to reflect the relative importance of features [102]. Such frameworks provide the

possibility of incorporating a wide range of scene understanding techniques into future visual representations, but at this stage have not been evaluated.

9.3 Conclusion

The work presented in this thesis establishes a new paradigm for prosthetic vision scene representation based on structural saliency. The results of the work indicate the significant potential for the application of structural scene analysis methods to create visual representations that can further improve functional outcomes for prosthetic vision users.

Bibliography

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision*, pages 1597–1604, 2009.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[3] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[4] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[5] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80. IEEE, 2010.

[6] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

[7] Lauren N Ayton, Peter J Blamey, Robyn H Guymer, Chi D Luu, David AX Nayagam, Nicholas C Sinclair, Mohit N Shivdasani, Jonathan Yeoh, Mark F McCombe, Robert J Briggs, et al. First-in-human trial of a novel suprachoroidal retinal prosthesis. *PloS One*, 9(12):e115239, 2014.

[8] Paul Bach-y Rita, Kurt A Kaczmarek, Mitchell E Tyler, and Jorge Garcia-Lara. Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note. *Journal of Rehabilitation Research and Development*, 35(4):427, 1998.

[9] Hernan Badino, Daniel Huber, Yongwoon Park, and Takeo Kanade. Fast and accurate computation of surface normals from range images. In *Proceedings of*

the IEEE International Conference on Robotics and Automation, pages 3084–3091. IEEE, 2011.

[10] Nick Barnes, Xuming He, Chris McCarthy, Lachlan Horne, Junae Kim, Adele Scott, and Paulette Lieby. The role of vision processing in prosthetic vision. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 308–311. IEEE, 2012.

[11] Nick Barnes, Adele F Scott, Paulette Lieby, Matthew A Petoe, Chris McCarthy, Ashley Stacey, Lauren N Ayton, Nicholas C Sinclair, Mohit N Shivdasani, Nigel H Lovell, et al. Vision function testing for a suprachoroidal retinal prosthesis: effects of image filtering. *Journal of Neural Engineering*, 13(3):036013, 2016.

[12] Nick Barnes, Adele F Scott, Ashley Stacey, Paulette Lieby, Chris McCarthy, Matt Petoe, Lauren N Ayton, Mohit Naresh Shivdasani, Nicholas C Sinclair, and Janine Walker. Vision processing with lanczos2 improves low vision test results in implanted visual prosthetic patients. *Investigative Ophthalmology & Visual Science*, 55(13):1802–1802, 2014.

[13] Nick Barnes, Janine Walker, David Feng, Chris D Mccarthy, Adele F Scott, Ashley Stacey, Rebecca Dengate, Lauren N Ayton, Matthew A Petoe, Nigel H Lovell, Penelope Allen, Anthony Burkitt, and Robyn H Guymer. First-in-human trial of a novel suprachoroidal retinal prosthesis. *PloS One*, under review.

[14] Olga Regina Pereira Bellon, Alexandre Ibrahim Direne, and Luciano Silva. Edge detection to guide range image segmentation by clustering techniques. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 725–729. IEEE, 1999.

[15] Olga RP Bellon and Luciano Silva. New improvements to range image segmentation by edge detection. *IEEE Signal Processing Letters*, 9(2):43–45, 2002.

[16] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389, 2015.

[17] Michael Beyeler, Geoffrey M Boynton, Ione Fine, and Ariel Rokem. pulse2percept: A python-based simulation framework for bionic vision. *bioRxiv*, page 148015, 2017.

[18] Johann Borenstein and Yoram Koren. The vector field histogram-fast obstacle avoidance for mobile robots. *IEEE Transactions on Robotics and Automation*, 7(3):278–288, 1991.

[19] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[20] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[21] Ali Borji, Dicky N. Sihite, and Laurent Itti. Salient object detection: A benchmark. In *Procedings of the IEEE European Conference on Computer Vision*, pages 414–429. 2012.

[22] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.

[23] Kichul Cha, Kenneth W Horch, and Richard A Normann. Mobility performance with a pixelized vision system. *Vision Research*, 32(7):1367–1372, 1992.

[24] Christel Chamaret, Sylvain Godeffroy, Patrick Lopez, and Olivier Le Meur. Adaptive 3d rendering based on region-of-interest. In *Proceedings of the SPIE Conference on Image, Sensors, and Applications*, pages 75240V–75240V, 2010.

[25] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 914–921. IEEE, 2011.

[26] Fangfang Chen, Congyan Lang, Songhe Feng, and Zehai Song. Depth information fused salient object detection. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 66:66–66:70. ACM, 2014.

[27] Hao Chen, YF Li, and Dan Su. Rgb-d salient object detection based on discriminative cross-modal transfer learning. *arXiv preprint arXiv:1703.00122*, 2017.

[28] Spencer C Chen, Gregg J Suaning, John W Morley, and Nigel H Lovell. Simulating prosthetic vision: Ii. measuring functional capacity. *Vision research*, 49(19):2329–2343, 2009.

[29] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[30] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1536, 2013.

[31] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–416, 2011.

[32] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, page 23, 2014.

[33] Changhyun Choi, Alexander JB Trevor, and Henrik I Christensen. Rgb-d edge detection and edge-based registration. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1568–1575. IEEE, 2013.

[34] Euisun Choi and Chulhee Lee. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709, 2003.

[35] Arridhana Ciptadi, Tucker Hermans, and James M Rehg. An in depth view of saliency. In *Proceedings of the British Machine Vision Conference*, pages 9–13, 2013.

[36] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[37] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149. IEEE, 2000.

[38] Gislin Dagnelie. Psychophysical evaluation for visual prosthesis. *Annual Review of Biomedical Engineering*, 10:339–368, 2008.

[39] Gislin Dagnelie, Pearse Keane, Venkata Narla, Liancheng Yang, James Weiland, and Mark Humayun. Real and virtual mobility performance in simulated prosthetic vision. *Journal of Neural Engineering*, 4(1):S92, 2007.

[40] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[41] Chloé De Balthasar, Sweta Patel, Arup Roy, Ricardo Freda, Scott Greenwald, Alan Horsager, Manjunatha Mahadevappa, Douglas Yanai, Matthew J McMahon, Mark S Humayun, et al. Factors affecting perceptual thresholds in epiretinal prostheses. *Investigative Ophthalmology & Visual Science*, 49(6):2303–2314, 2008.

[42] Karthik Desingh, Madhava Krishna K, Deepu Rajan, and CV Jawahar. Depth really matters: improving visual salient region detection with depth. In *Proceedings of the British Machine Vision Conference*, 2013.

[43] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1848, 2013.

[44] Jason A Dowling and Anthony J Maeder. Mobility enhancement and assessment for a visual prosthesis. International Society for Optical Engineering, 2004.

[45] Jason A Dowling, Anthony J Maeder, and Wageeh W Boles. A pda based artificial human vision simulator. 2005.

[46] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. Saliency detection for stereoscopic images. *IEEE Transactions on Image Processing*, 23(6):2625–2636, 2014.

[47] Olivier Faugeras, Joe Mundy, Narendra Ahuja, Charles Dyer, Alex Pentland, Ramesh Jain, Katsushi Ikeuchi, and Kevin Bowyer. Why aspect graphs are not (yet) practical for computer vision. *CVGIP: Image Understanding*, 55(2):212–218, 1992.

[48] Simone Frintrop, Andreas Nüchter, and Hartmut Surmann. Visual attention for object recognition in spatial 3d data. In *International Workshop on Attention and Performance in Computer Vision*, pages 168–182. 2004.

[49] Simone Frintrop, Andreas Nuchter, Hartmut Surmann, and Joachim Hertzberg. Saliency-based object recognition in 3d data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2167–2172. IEEE, 2004.

[50] Takashi Fujikado, Motohiro Kamei, Hirokazu Sakaguchi, Hiroyuki Kanda, Takeshi Morimoto, Yasushi Ikuno, Kentaro Nishida, Haruhiko Kishima, Tomoyuki Maruo, Kunihiko Konoma, et al. Testing of semichronically implanted retinal prosthesis by suprachoroidal-transretinal stimulation in patients with retinitis pigmentosa. *Investigative Ophthalmology & Visual Science*, 52(7):4726–4733, 2011.

[51] Yaroslav Ganin and Victor Lempitsky. N^ 4-fields: Neural network nearest neighbor fields for image transforms. In *Proceedings of the IEEE Asian Conference on Computer Vision*, pages 536–551. Springer, 2014.

[52] Carl Friedrich Gauss. *General investigations of curved surfaces of 1827 and 1825*. The Princeton University Library, 1902.

[53] Guido Gerig, Olaf Kubler, Ron Kikinis, and Ferenc A Jolesz. Nonlinear anisotropic filtering of mri data. *IEEE Transactions on medical imaging*, 11(2):221–232, 1992.

[54] James R Golden, Cordelia Erickson-Davis, Nicolas P Cottaris, Nikhil Parthasarathy, Fred Rieke, David H Brainard, Brian A Wandell, and EJ Chichilnisky. Simulation of visual perception and learning with a retinal prosthesis. *bioRxiv*, page 206409, 2017.

[55] François Goudail, Philippe Réfrégier, and Guillaume Delyon. Bhattacharyya distance as a contrast parameter for statistical processing of noisy optical images. *Journal of the Optical Society of America*, 21(7):1231–1240, 2004.

[56] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for rgb-d image via saliency evolution. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2016.

[57] Jingfan Guo, Tongwei Ren, Jia Bei, and Yujin Zhu. Salient object detection in rgb-d image based on saliency fusion and propagation. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, page 59. ACM, 2015.

[58] Jingfan Guo, Tongwei Ren, Jia Bei, and Yujin Zhu. Salient object detection in RGB-D image based on saliency fusion and propagation. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 59:1–59:5, 2015.

[59] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[60] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[61] Gideon Guy and Gérard Medioni. Inferring global pereeptual contours from local features. *International Journal of Computer Vision*, 20(1):113–133, 1996.

[62] Jasmine S Hayes, Vivian T Yin, Duke Piyathaisere, James D Weiland, Mark S Humayun, and Gislin Dagnelie. Visually guided performance of simple tasks using simulated prosthetic vision. *Artificial Organs*, 27(11):1016–1028, 2003.

[63] Xuming He, Junae Kim, and Nick Barnes. An face-based visual fixation system for prosthetic vision. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2981–2984. IEEE, 2012.

[64] Ian P Howard and Brian J Rogers. *Binocular vision and stereopsis*. Oxford University Press, USA, 1995.

[65] Mark S Humayun, Jessy D Dorn, Lyndon Da Cruz, Gislin Dagnelie, José-Alain Sahel, Paulo E Stanga, Artur V Cideciyan, Jacque L Duncan, Dean Eliott, Eugene Filley, et al. Interim results from the international trial of second sight's visual prosthesis. *Ophthalmology*, 119(4):779–788, 2012.

[66] Mark S Humayun, James D Weiland, Gildo Y Fujii, Robert Greenberg, Richard Williamson, Jim Little, Brian Mech, Valerie Cimmarusti, Gretchen Van Boemel,

Gislin Dagnelie, et al. Visual perception in a blind subject with a chronic microelectronic retinal prosthesis. *Vision Research*, 43(24):2573–2581, 2003.

[67] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[68] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[69] Dorothea Jameson and Leo M Hurvich. Theory of brightness and color contrast in human vision. *Vision Research*, 4(1):135–154, 1964.

[70] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2013.

[71] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2013.

[72] Xiaoyi Jiang and Horst Bunke. Edge detection in range images based on scan line approximation. *Computer Vision and Image Understanding*, 73(2):183–199, 1999.

[73] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1115–1119. IEEE, 2014.

[74] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. *Proceedings of the IEEE International Conference on Image Processing*, 2014.

[75] KA Kaczmarek. The tongue display unit (tdu) for electrotactile spatiotemporal pattern presentation. *Scientia Iranica*, 18(6):1476–1485, 2011.

[76] Kurt A Kaczmarek, John G Webster, Paul Bach-y Rita, and Willis J Tompkins. Electrotactile and vibrotactile displays for sensory substitution systems. *IEEE Transactions on Biomedical Engineering*, 38(1):1–16, 1991.

[77] Pattaraporn Khuwuthyakorn, Antonio Robles-Kelly, and Jun Zhou. Object of interest detection by saliency learning. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 636–649. Springer, 2010.

[78] Josef Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42, 1983.

[79] Jyri J Kivinen, Christopher KI Williams, Nicolas Heess, et al. Visual boundary prediction: A deep neural prediction network and quality dissection. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 1, page 9, 2014.

[80] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2214–2219. IEEE, 2011.

[81] Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. *Proceedings of the International Conference on Learning Representations*, 2016.

[82] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: influence of depth cues on visual saliency. In *ECCV*, pages 101–115. 2012.

[83] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–668, 2016.

[84] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Eld-net: An efficient deep learning architecture for accurate saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[85] Alex Levinshtein, Cristian Sminchisescu, and Sven Dickinson. Optimal contour closure by superpixel grouping. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 480–493. Springer, 2010.

[86] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

[87] Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Contextual hypergraph modeling for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3328–3335, 2013.

[88] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[89] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[90] Paulette Lieby, Adele Scott, Nick Barnes, Ashley Stacey, Lauren Ayton, and Janine Walker. Evaluating lanczos2 image filtering for visual acuity in simulated

prosthetic vision. *Investigative Ophthalmology & Visual Science*, 54(15):1049–1049, 2013.

[91] Wei-Yang Lin, Pei-Chen Wu, and Bo-Rong Chen. Image retargeting using depth enhanced saliency. *Proceedings of the International Conference on 3D Systems and Applications*, pages 1–4, 2013.

[92] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[93] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[94] Yu Liu and Michael S Lew. Learning relaxed deep supervision for better edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 231–240, 2016.

[95] David Lowe. *Perceptual organization and visual recognition*, volume 5. Springer Science & Business Media, 2012.

[96] Song Lu, Vijay Mahadevan, and Nuno Vasconcelos. Learning optimal seeds for diffusion-based salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797, 2014.

[97] Yanyu Lu, Han Kan, Jie Liu, Jing Wang, Chen Tao, Yao Chen, Qiushi Ren, Jie Hu, and Xinyu Chai. Optimizing chinese character displays improves recognition and reading performance of simulated irregular phosphene mapsoptimizing chinese character displays. *Investigative Ophthalmology & Visual Science*, 54(4):2918–2926, 2013.

[98] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 580–596. Springer, 2016.

[99] Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1139–1146, 2013.

[100] James A Marron and Ian L Bailey. Visual factors and orientation-mobility performance. *Optometry and Vision Science*, 59(5):413–426, 1982.

[101] Chris McCarthy and Nick Barnes. Surface extraction from iso-disparity contours. In *Asian Conference on Computer Vision*, pages 410–421. Springer, 2010.

[102] Chris McCarthy and Nick Barnes. Importance weighted image enhancement for prosthetic vision: An augmentation framework. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 45–51. IEEE, 2014.

[103] Chris McCarthy, Nick Barnes, and Paulette Lieby. Ground surface segmentation for navigation with a low resolution visual prosthesis. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4457–4460. IEEE, 2011.

[104] Chris McCarthy, Janine G Walker, Paulette Lieby, Adele Scott, and Nick Barnes. Mobility and low contrast trip hazard avoidance using augmented depth. *Journal of Neural Engineering*, 12(1):016003, 2014.

[105] Paria Mehrani and Olga Veksler. Saliency segmentation based on learning and graph cut refinement. In *Proceedings of the British Machine Vision Conference*, pages 1–12, 2010.

[106] Paria Mehrani and Olga Veksler. Saliency segmentation based on learning and graph cut refinement. In *Proceedings of the British Machine Vision Conference*, pages 1–12, 2010.

[107] Yansheng Ming, Hongdong Li, and Xuming He. Winding number constrained contour detection. *IEEE Transactions on Image Processing*, 24(1):68–79, 2015.

[108] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461, 2012.

[109] Bruno A Olshausen, Charles H Anderson, and David C Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993.

[110] Nabil Ouerhani and Heinz Hügli. Computing visual attention from scene depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–378, 2000.

[111] N Parikh, L Itti, M Humayun, and J Weiland. Performance of visually guided tasks using simulated prosthetic vision and saliency-based cues. *Journal of Neural Engineering*, 10(2):026017, 2013.

[112] Neha Parikh, Laurent Itti, and James Weiland. Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering*, 7(1):016006, 2010.

[113] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Proceedings of the IEEE European Conference on Computer Vision*, 2014.

[114] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 92–109. 2014.

[115] Serge Picaud and José-Alain Sahel. Retinal prostheses: Clinical results and future challenges. *Comptes Rendus Biologies*, 337(3):214–222, 2014.

[116] Ekaterina Potapova, Michael Zillich, and Markus Vincze. Learning what matters: combining probabilistic models of 2d and 3d saliency cues. In *Computer Vision Systems*, pages 132–142. 2011.

[117] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[118] Vilayanur S Ramachandran. Perception of shape from shading. *Nature*, 331(6152):163–166, 1988.

[119] Ramesh Raskar, Kar-Han Tan, Rogerio Feris, Jingyi Yu, and Matthew Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. In *ACM Transactions on Graphics*, volume 23, pages 679–688. ACM, 2004.

[120] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Yang. Exploiting global priors for RGB-D saliency detection. In *The IEEE Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32, 2015.

[121] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for rgb-d saliency detection. In *Proceedings of the IEEE Conference on Computer VIsion and Pattern Recognition-Workshops*, pages 25–32, 2015.

[122] Constantino Carlos Reyes-Aldasoro and Abhir Bhalerao. The bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5):812–826, 2006.

[123] Joseph F Rizzo III. Update on retinal prosthetic research: the boston retinal implant project. *Journal of Neuro-ophthalmology*, 31(2):160–168, 2011.

[124] Eliana Sampaio, Stéphane Maris, and Paul Bach-y Rita. Brain plasticity:'visual'acuity of blind persons via the tongue. *Brain Research*, 908(2):204–207, 2001.

[125] Angel Domingo Sappa. Unsupervised contour closure algorithm for range image edge-based segmentation. *IEEE Transactions on Image Processing*, 15(2):377–384, 2006.

[126] Angel Domingo Sappa and Michel Devy. Fast range image segmentation by an edge detection strategy. In *Proceedings of the IEEE International Conference on 3-D Digital Imaging and Modeling*, pages 292–299. IEEE, 2001.

[127] Henrik Schäfer, Frank Lenzen, and Christoph S Garbe. Depth and intensity based edge detection in time-of-flight images. In *Proceedings of the International Conference on 3DTV-Conference*, pages 111–118. IEEE, 2013.

[128] Bjoern Schwander. Early health economic evaluation of the future potential of next generation artificial vision systems for treating blindness in germany. *Health Economics Review*, 4(1):27, 2014.

[129] Alexander G Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2815–2822. IEEE, 2012.

[130] RM Shapley and DJ Tolhurst. Edge detectors in human vision. *The Journal of Physiology*, 229(1):165–183, 1973.

[131] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2015.

[132] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[133] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning rgbd salient object detection using background enclosure, depth contrast, and top-down features. In *IEEE International Conference on Computer Vision-Workshops*, 2017.

[134] Shraga Shoval, Iwan Ulrich, and Johann Borenstein. Navbelt and the guidecane [obstacle-avoidance systems for the blind and visually impaired]. *IEEE Robotics & Automation Magazine*, 10(1):9–20, 2003.

[135] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[136] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, and Cong Bai. Saliency detection for rgbd images. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, page 72. ACM, 2015.

[137] Grace P Soong, Jan E Lovie-Kitchin, and Brian Brown. Does mobility performance of visually impaired adults improve immediately after orientation and mobility training? *Optometry & Vision Science*, 78(9):657–666, 2001.

[138] Katarina Stingl, Karl Ulrich Bartz-Schmidt, Dorothea Besch, Angelika Braun, Anna Bruckmann, Florian Gekeler, Udo Greppmaier, Stephanie Hipp, Gernot Hörtdörfer, Christoph Kernstock, et al. Artificial vision with wirelessly powered

subretinal electronic implant alpha-ims. In *Proceedings of the Royal Society of London B: Biological Sciences*, volume 280, page 20130077. The Royal Society, 2013.

[139] Katarina Stingl, Ruth Schippert, Karl U Bartz-Schmidt, Dorothea Besch, Charles L Cottriall, Thomas L Edwards, Florian Gekeler, Udo Greppmaier, Katja Kiel, Assen Koitschev, et al. Interim results of a multicenter trial with the new electronic subretinal implant alpha ams in 15 patients blind from inherited retinal degenerations. *Frontiers in Neuroscience*, 11:445, 2017.

[140] H Christiaan Stronks and Gislin Dagnelie. Phosphene mapping techniques for visual prostheses. In *Visual Prosthetics*, pages 367–383. Springer, 2011.

[141] H Christiaan Stronks, Daniel J Parker, and Nick Barnes. Vibrotactile spatial acuity and intensity discrimination on the lower back using coin motors. *IEEE Transactions on Haptics*, 9(4):446–454, 2016.

[142] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[143] Shuai Tang, Xiaoyu Wang, Xutao Lv, Tony X Han, James Keller, Zhihai He, Marjorie Skubic, and Shihong Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *Proceedings of the IEEE Asian Conference on Computer Vision*, pages 525–538. Springer, 2012.

[144] Yanlong Tang, Ruofeng Tong, Min Tang, and Yun Zhang. Depth incorporating with color improves salient object detection. *The Visual Computer*, 32(1):111–121, 2016.

[145] Mary E Tinetti. Performance-oriented assessment of mobility problems in elderly patients. *Journal of the American Geriatrics Society*, 34(2):119–126, 1986.

[146] Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7):4, 2009.

[147] TTHUR15000. "http://mmcheng.net/gsal/".

[148] Ken Turkowski. Filters for common resampling-tasks. *Graphics Gems*, pages 147–165, 1990.

[149] Shimon Ullman and Amnon Sha'ashua. Structural saliency: The detection of globally salient structures using a locally connected network. 1988.

[150] Gianni Virgili and Gary Rubin. Orientation and mobility training for adults with low vision. *Cochrane Database Syst Rev*, 3, 2006.

[151] Junle Wang, Matthieu Perreira DaSilva, Patrick LeCallet, and Vincent Ricordel. Computational model of stereoscopic 3d visual saliency. *IEEE Transactions on Image Processing*, 22(6):2151–2165, 2013.

[152] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2015.

[153] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. *Proceedings of the IEEE European Conference on Computer Vision*, pages 29–42, 2012.

[154] Bernard Widrow. A study of rough amplitude quantization by means of nyquist sampling theory. *IRE Transactions on Circuit Theory*, 3(4):266–276, 1956.

[155] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[156] Ren Xiaofeng and Liefeng Bo. Discriminatively trained sparse code gradients for contour detection. In *Advances in Neural Information Processing Systems*, pages 584–592, 2012.

[157] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[158] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[159] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–202, 2016.

[160] Naokazu Yokoya and Martin D Levine. Range image segmentation based on differential geometry: A hybrid approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):643–649, 1989.

[161] Yun Zhang, Gangyi Jiang, Mei Yu, and Ken Chen. Stereoscopic visual attention model for 3d video. In *Advances in Multimedia Modeling*, pages 314–324. 2010.

[162] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[163] Ying Zhao, Yanyu Lu, Yukun Tian, Liming Li, Qiushi Ren, and Xinyu Chai. Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision. *Information Sciences*, 180(16):2915–2924, 2010.

[164] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014.

[165] Eberhart Zrenner, Karl Ulrich Bartz-Schmidt, Heval Benav, Dorothea Besch, Anna Bruckmann, Veit-Peter Gabel, Florian Gekeler, Udo Greppmaier, Alex Harscher, Steffen Kibbel, et al. Subretinal electronic chips allow blind patients to read letters and combine them to words. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1711):1489–1497, 2011.